National Library of the Czech Republic

Facilitating Access to Cultural Heritage Content in Czechia

INTERMI project

Marie Balíková
Miroslav Kunt
Jana Šubová
Nadežda Andrejčíková

Prague 2015

KATALOGIZACE V KNIZE - NÁRODNÍ KNIHOVNA ČR

Balíková, Marie
Facilitating access to cultural heritage content in Czechia : INTERMI project / Marie Balíková, Miroslav Kunt, Jana Šubová, Nadežda Andrejčíková. -- First edition. -- Prague : National Library of the Czech Republic, 2015. -- 1 online zdroj
ISBN 978-80-7050-662-2 (online)

026/027+069+930.25 * 930.85 * [024.5+659.2]-021.161 * 004.8:111.1 * 004.41.045 * 004.652.4/.5 * 005.591-027.542 * (437.3)
- memory institutions -- Czechia
- cultural heritage
- accessibility of information
- ontology (informatics)
- object-oriented methods
- object-relational database systems
- national programs and projects – Czechia
- collective monographs

- Interoperabilita v paměťových institucích (projekt)
- paměťové instituce -- Česko
- kulturní dědictví
- zpřístupňování informací
- ontologie (informatika)
- objektově orientované metody
- objektově-relační databázové systémy
- národní programy a projekty -- Česko
- kolektivní monografie

020 - Library and information sciences [12]
02 - Knihovnictví. Informatika [12]

This publication is the outcome of the project “Interoperability in memory institutions (INTERMI)”. The project was funded in 2011–2015 by Ministry of Culture within the framework of NAKI program (DF11P01OVV023).

Investigators of the project: Mgr. Marie Balíková, National Library of the Czech Republic; Ing. Miroslav Kunt, National Archives

Reviewer of the publication: PhDr. Barbora Drobíková, Ph.D.

More information: http://www.interpi.cz

First edition

ISBN 978-80-7050-662-2

Content

Preface

It is said that collections of memory institutions form the heart of national cultural, industrial and natural heritage. Now with more and more information about national cultural heritage content becoming available thanks to the mass digitization of information resources, there is an increasing demand for global search across memory institutions collections. These institutions – libraries, archives and museums – desire to share the information about the cultural heritage content that are responsible for, and to offer it to the wider public. They investigate the ways how to do that. Libraries, archives and museums see now a great opportunity by opening up their collections to the general public using Semantic Web technologies. This requires the change of thinking, the courage to critically evaluate the standards, rules, procedures and practices applied now by describing library, archival and museums materials and to discover new approaches, new possibilities. We tried it within the project INTERMI. The aim of the project is to create an ontology and a conceptual model of knowledge which meets the requirements and needs of all memory institutions users and professionals as well. We offer them an ontology for cultural heritage content based on semantic interoperability and Semantic Web technologies. The project also brings a new paradigm of data processing based on object-oriented approach which is focused on processing of entities (classes) and on complex relations among them. This book presents the results achieved within the five-year project. They were favourably received and highly evaluated internationally, especially at the last IFLA World Library and Information Congress in Cape Town 2015; we believe that they will be useful and beneficial as a platform for further research.

List of abbreviations

.NET	software framework developed by Microsoft
AACR, AACR2	Anglo-American Cataloguing Rules
AAT	The Art & Architecture Thesaurus
AMC	Archival and Manuscripts Control
ANSI/NISO	American National Standards Institute/National Information Standards Organization
APPM	Archives, Personal Papers, and Manuscripts
ARIA	Accessible Rich Internet Applications
BIBFRAME	Bibliographic Framework (BIBFRAME Initiative, BIBFRAME Model, BIBFRAME Vocabulary)
C++	general-purpose programming language
CCO	Cataloging Cultural Object
CDWA	Categories for the Description of Works of Art
CES	Central Registry of Museum-type Collections
CIDOC CR	CIDOC Conceptual Reference Model
CIDOC	International Committee for Documentation
CITeM	Centrum pro informační technologie v muzejnictví
COM	Component Object Model
CONA	Cultural Objects Name Authority Online
CSP	Caché Server Pages
CSV	Comma-separated values
DA:CS	Describing Archives: A Content Standard
DACS	Describing Archives: A Content Standard (national application of ISAD (G))
DDC	Dewey Decimal Classification
DOC	word processing file format
DTD	Document Type Definition
EAD	Encoded Archival Description
EAD-CPF	Encoded Archival Context – Corporate bodies, Persons and Families
EGAD	Experts Group on Archival Description
EJB	Enterprise Java Beans
EuroVoc	multilingual, multidisciplinary thesaurus covering the activities of the EU
FRAD	Functional Requirements for Authority Data
FRBR ($FRBR_{ER}$)	Functional Requirements for Bibliographic Records
$FRBR_{OO}$	FRBR – object oriented
FRSAD	Functional Requirements for Subject Authority Data
HTTP	Hypertext Transfer Protocol
ICA	International Council on Archives
ICOM	International council of museums
ICT	information and communication technologies

ID	identifier
IFLA	International Federation of Library Associations and Institutions
ILS	Integrated Library Systems
ISAAR (CPF)	International Standard Archival Authority Record for Corporate Bodies, Persons and Families
ISAD (G)	General International Standard Archival Description
ISBD	International Standard Bibliographic Description
ISDF	International Standard for Describing Functions
ISDIAH	International Standard for Describing Institutions with Archival Holdings
ISO	International Organization for Standardization
LC	Library of Congress
LCC	Library of Congress Classification
LCNAF	Library of Congress Name Authority Files
LCSH	Library of Congress Subject Headings
LESS	stylesheet language
LIDO	Lightweight Information Describing Objects
LOD	Linked Open Data
MARC, MARC21	MAchine Readable Cataloging
MVC/MVVM	Model-View-Controller / Model View ViewModel
NAKI	Národní kulturní identita (National culture identity)
NLCR	National Library of the Czech Rupublic (Národní knihovna České republiky)
OAI	Open Archives Initiative
OCLC	Online Computer Library Center
ODMG	Object Database Management Group
OPAC	Online Public Access Catalogue
OWL	Web Ontology Language
PDF	Portable Document Format
PeVa	Archive Groups in the Czech Republic
PSH	Polythematic Structured Subject Heading System (Polytematický strukturovaný heslář)
PUC	Permanent UNIMARC Committee
RDA	Resource Description and Access
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RDSMS	Relational Data Stream Management System
REST	Representational State Transfer
RPC-style	Remote Procedure Call
SAA	Society of American Archivists
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organization System
SOAP	Simple Object Access protocol
SPARQL	SPARQL Protocol and RDF Query Language

SQL	Structured Query Language
SRU	Search/Retrieve via URL
SRW	Search Retrieve Web Service
TCP/IP	Transmission Control Protocol/Internet Protocol
TGN	The Getty Thesaurus of Geographic Names
UDC	Universal Decimal Classification
UDDI	Universal Description, Discovery and Integration
ULAN	The Union List of Artist Names
UNIMARC	Universal MARC format
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VRA core	Visual Resources Association Core
W3C	World Wide Web Consortium
WSDL	Web Service Description Language
WWW	World Wide Web
XML	eXtensible Markup Language
XSL	eXtensible Stylesheet Language

Introduction

The aim of the project is to create scientific and technological infrastructure to support the processing, sharing and use of the cultural content in the form of metadata about information objects preserved in memory institutions. The information objects represent objects of the real world, e.g. persons, institutions, activities, three-dimensional objects (i.e. artistic and technical objects, objects of inanimate nature), events and performances, artistic and other achievements as well as their textual, visual and sound representations (digital included), which are subject to the activity of memory institutions.

A basic pre-requisite for identification of the information objects are both subject and name authorities, in the extent defined by the needs of libraries. The aim of this project is to extend the function of authorities by adding related information objects: in this way, their information and documentation potentiality will increase so that they would be able not only to meet the requirements of different types of memory institutions, but to create much larger basis for semantic interoperability of cultural content of all of them.

One of the main results of the project will be a tool for user-friendly and differentiated acquisition and accessing of information about these information objects. This will also create a basic pre-requisite for the access and share of the cultural content across the memory institutions ensuring semantic interoperability at the conceptual level. All memory institutions as well as general public audience are invited to use the proposed system.

The project is complementary to the projects such as the National digital library, the National digital archive. The proposed project builds on earlier research goals, but in addition it aims for a comprehensive construction of an infrastructure for building a knowledge based model of the cultural heritage and for the opening of its use in the form of working pilot operation.

From the very beginning of the project it was clear, that the development and creation of scientific and technological infrastructure of a common base for cultural heritage content in Czechia would be a very demanding process from several reasons. The first reason is that archives, museums, and galleries in Czechia remained relatively closed to the general public until the year 2000. In addition, the funds of archives and museums are unique, therefore these institutions did not feel the need to share them broadly and to develop common standards, rules and systems for their user-friendly access. These communities started to develop rules and standards at the beginning of the 21st century, e.g. Basic rules for archival processing in Czechia were published in 2013. Therefore, in the first phase of the project, we dealt with the analysis of the best standards and rules used in foreign countries and possibilities of their application in our environment.

Another very important reason is that the information technology environment is rather complex and volatile. It is generally known that memory institutions produce information of high quality that comply with rules and standards;

however this information is unavailable on the Web, as being part of the so called Invisible Web. In addition, the research studies based on user statistics show that users are moving away from library websites and catalogues when they look for information. They prefer big, strong, freely available search engines like Google, even if the information published there can be misleading and unverified. They consider search options offered by library websites and catalogues outdated and cumbersome. Professionals from all communities (information and communication technologies, libraries, archives, museums...) are aware of the issues and try to find appropriate solutions. Interesting and successful, but challenging solution (above all for library community) come from linking Open Data community: they offer Linked Open Data (LOD) publication model. It is clear that current data produced by the institutions according to standards must be changed and adapted, and the standards must be revised and adapted as well, and in some cases new standards should be created (MARC format, BIBFRAME).

Also, the library and the museum community have realized the seriousness of the situation and the urgency to deal with it. In 1990s, library and museum communities decided to try to overcome the heterogeneity of information produced and published on the Web by applying conceptual modelling techniques in defining data models which could better represent its data. The first Entity-Relationship Models were developed and discussed (e.g. FRBR, FRAD, FRSAD). Museum community preferred to apply Object-Oriented Models (e.g. CIDOC CRM). Recent developments proved that the object-oriented approach would be more convenient for the bibliographic universe as well ($FRBR_{OO}$).

1 Memory institutions in Czechia

The mission of individual memory institutions in Czechia is different comparing with the situation in e.g. English speaking countries. While the mission of libraries has been the same: the libraries have been responsible for developing policies on the acquisition, documentation, conservation, research, and communication, it means for the open information access to mostly textual/printed heritage (documents), the mission of other memory institutions in Czechia was different. Museums and art galleries have been responsible for collecting, documenting (scientific research included), preserving and displaying art collections and museum objects/artefacts via permanent and temporary exhibitions, archives for collection, documentation, and preservation of information as evidence of the creator‘s activities.

By the tradition, the archives and museums in Czechia have been considered as research institutions and have been resistant towards the openness to the public audience: archivists, curators and small group of researchers are allowed to decide what and how much to make available to the public and what to keep hidden in their reserves.

1.1 Libraries

Libraries have well-developed services, library systems, rules and metadata schemas, advanced search techniques; they have been able to share controlled vocabularies, authority files, as well as to provide easy and intuitive access to information for humans.

Title	**Simpsonovi a filozofie** : Homer myslitel / editoři William Irwin, Mark T. Conard, Aeon J. Skoble ; [z anglického originálu ... přeložili Jan Žlábek ... et al.]		
Imprint	V Praze : XYZ, 2010		
Physical descr.	443 s. ; 21 cm		
HOLDINGS			
ISBN	978-80-7388-216-7 (váz.)		
Variant title	Obálkový a hřbetní název: **Simpsonovi & filozofie**		
Translation of	Simpsons and philosophy		
Other author	Irwin, William, 1970-		
	Conard, Mark T., 1965-		
	Skoble, Aeon J., 1964-		
Note	Obsahuje bibliografické odkazy		
		Conspectus	791 : Film. Cirkus. Lidová zábava
Subject heading	**Simpsonovi** (TV seriál)	UDC	791.242.097
Subject heading	televizní seriály -- Spojené státy americké		791.228
	animované filmy -- Spojené státy americké -- 1989-		791-5
	filmové postavy		17.022.1
	morální hodnoty		7.011
	masová kultura -- filozofické aspekty		151
Chronolog. term	* od 1989		(73)
Genre/form	* eseje		(0.82-4)
		Subject head.-eng.	Simpsons (television program)
		Subject head.-eng.	* television serials -- United States * animated film -- United States -- 1989- * film characters * moral values * popular culture -- philosophical aspects
		Genre/form-eng.	* essays

Figure 1: Example of access to information resource applied in libraries: bibliographic record

Libraries have a long history of using shared international standards and tools within their community. A common descriptive standard Anglo-American Cataloguing

Rules (AACR) was published in 1967 and used to produce standardized bibliographic and authority records, which were used and shared by national bibliographic agencies and libraries at both the national and international level. The standard was translated into Czech language in 1995 and then became part of cataloguing practice in Czechia. Since 2015, the content standard AACR has been replaced by its successor Resource Description and Access (RDA) rules.

When adopting common descriptive and other rules in bibliographic area, libraries can share not only bibliographic records, but controlled vocabularies, authority files, and thesauri as well. The main sources of authorised access points for names in English speaking countries have been the Library of Congress Name Authority Files (LCNAF). In subject analysis area, the Library of Congress Subject Headings (LCSH), LCC, DDC, the Thesaurus for Graphic Materials (TGM I & II) have been adopted.

The exchange of standardized bibliographic and authority data among libraries around the world was facilitated by the development and creation of MAchine Readable Cataloging format (MARC, ISO 2709) in the 1960s. Libraries can participate in shared or "copy" cataloguing process, and integrate records in union catalogues. Since advent of Internet in 1990s, libraries have been able to apply network protocols such as Z39.50/SRU (Search/Retrieve via URL) and SRW (Search Retrieve Web Service), and OAI.

The situation of libraries in non-English speaking countries was quite different: they had to face a lot of issues to be solved, especially in the subject analysis area. It was not possible to accept and apply the data values standards published in English language, e.g. controlled vocabularies, classification schemas without their adaptations, because of their biased approach and inability to meet their users' needs or to cover and respect all cultural and social differences. The library communities in non-English speaking countries were forced at least to translate or to develop their own controlled vocabularies, authority files, and classification systems (e.g. Universal Decimal Classification – Mezinárodní desetinné třídění) into the language/s of the country, and share them at the national level.

AUT - Full View of Record

Choose format: Standard -- MARC

Record 1 out of 1

Control no.	**unn2010590021**
Heading	**Simpsonovi (TV seriál)**
Seen from	Simpsons (**TV seriál**)
Public gen. note	Televizní **seriál**, animovaný, komedie: USA, 1989.
Source	**Simpsonovi** a filozofie (editor W.Irwin)
	www(FOX Broadcasting Company)
	www(Česko-Slovenská filmová databáze)
More info	Oficiální stránka korporace: http://www.thesimpsons.com/

Figure 2: Example of access to information resource applied in libraries: authority record

1.2 Archives

In the archives, the information resources are arranged according to the provenance principle of archive fonds. This means that the collections of historical records have been created and maintained for specific purposes such as administrative, business or legal needs, and they have been preserved for their long-term archival value. Everyone has access to information resources kept in the archives with respect to the valid legal norms: study records, their copies, or finding aids can be studied either in the research rooms of archives or via exhibitions, lectures or educational activities (Wanner, 2005, p. 2).

Large collections of archives contain the unique primary source materials that are described mostly at collection level using registers or inventories called finding aids. Traditionally, finding aids are locally developed, in hand- or typewritten form; they represent home-grown tools used in one repository, therefore it was very difficult to share them broadly.

After the MARC format was developed and applied, the archives in English speaking countries tried to share archival information by cataloguing collections in library systems. Afterwards, they created a MARC Format defined for Archival and Manuscripts Control (AMC). To facilitate the description of archival materials within MARC bibliographic systems, the archives in Canada and U.S. created their own content standard called Archives, Personal Papers, and Manuscripts (APPM) based on AACR rules. With advent of Internet and later, in the 1990s, of the Web, it was possible to make the finding aids available online, although it was not possible to express the complex hierarchical relationships using this flat data structure format (MARC). Therefore in 1999, MARC format was replaced by Encoded Archival Description (EAD) format which permits sharing the complex hierarchical relationships inherent in archival collections. Since then, EAD has become a common standard for sharing archival finding aids electronically. As a data format, the Standard Generalized Markup Language (SGML) and XML (2002) were used. In 2004, a new descriptive (content) standard emerged: Describing Archives: A Content Standard (DA:CS), so that the finding aids could be shared broadly online.

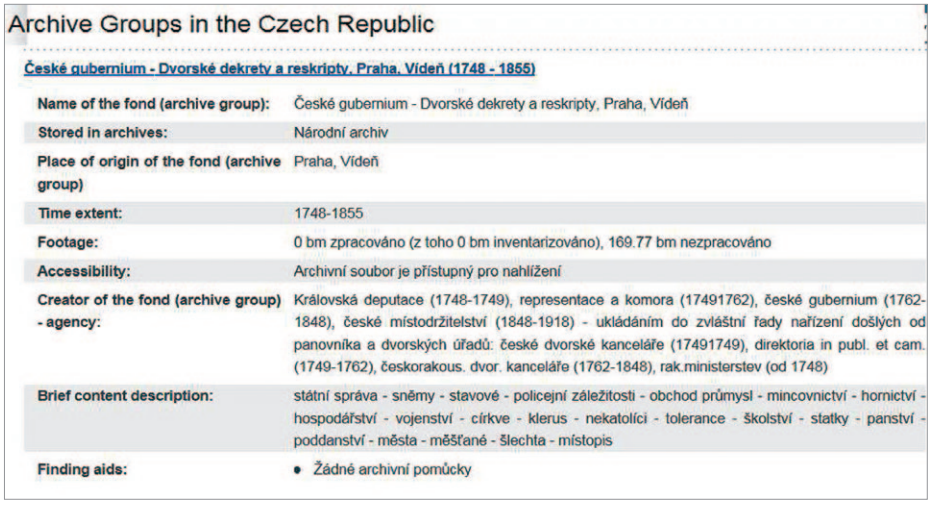
As was mentioned, the archives in Czechia have not been forced to share their information broadly, therefore a wide range of local systems and home-grown indexes, registers and finding aids were applied. The terminology used in indexes and registers was limited to local (one collection, one institution, one system) practice.

In 1990s, after the advent of Internet and the Web, the online databases Archive Groups in the Czech Republic and Badatelna.eu were launched.

1.2.1 Archive Groups in the Czech Republic (PeVa)

The database contains documents of permanent value deposited in the archives (historical records and documents) from archive fonds and archive collections. They represent so called basic registered units in archives. The database consists

of entries with name of the fond, name of the archives wherein the relevant archives fond (archive group) is deposited, footage of the fond, place(es) of origin of the fond, name of the originator in the original language and wording, time extent of the fond and finding aids available at the archives where the fond is stored.



Archive Groups in the Czech Republic

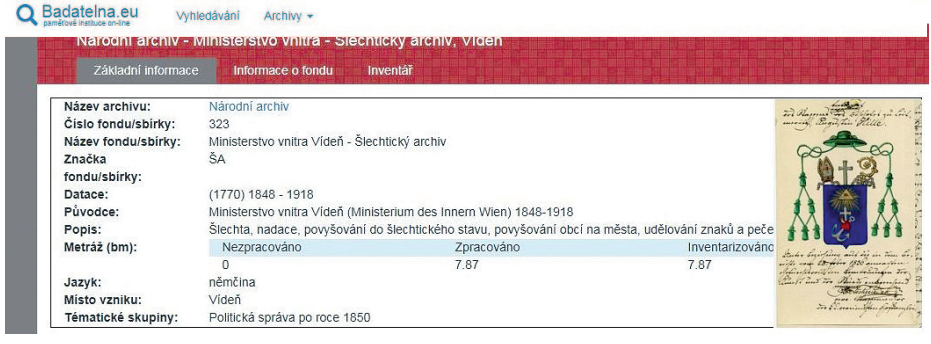
České gubernium - Dvorské dekrety a reskripty, Praha, Vídeň (1748 - 1855)

Name of the fond (archive group):	České gubernium - Dvorské dekrety a reskripty, Praha, Vídeň
Stored in archives:	Národní archiv
Place of origin of the fond (archive group)	Praha, Vídeň
Time extent:	1748-1855
Footage:	0 bm zpracováno (z toho 0 bm inventarizováno), 169.77 bm nezpracováno
Accessibility:	Archivní soubor je přístupný pro nahlížení
Creator of the fond (archive group) - agency:	Královská deputace (1748-1749), representace a komora (17491762), české gubernium (1762-1848), české místodržitelství (1848-1918) - ukládáním do zvláštní řady nařízení došlých od panovníka a dvorských úřadů: české dvorské kanceláře (17491749), direktoria in publ. et cam. (1749-1762), českorakous. dvor. kanceláře (1762-1848), rak.ministerstev (od 1748)
Brief content description:	státní správa - sněmy - stavové - policejní záležitosti - obchod průmysl - mincovnictví - hornictví - hospodářství - vojenství - církve - klerus - nekatolíci - tolerance - školství - statky - panství - poddanství - města - měšťané - šlechta - místopis
Finding aids:	• Žádné archivní pomůcky

Figure 3: Example of access to information resource applied database PeVa

1.2.2 Badatelna.eu portal

The Badatelna.eu portal gives users information about historical documents preserved by the Czech archives and enables displaying of digitized finding aids.



Badatelna.eu

Vyhledávání Archivy

Národní archiv - Ministerstvo vnitra - Šlechtický archiv, Vídeň

Základní informace | Informace o fondu | Inventář

Název archivu:	Národní archiv
Číslo fondu/sbírky:	323
Název fondu/sbírky:	Ministerstvo vnitra Vídeň - Šlechtický archiv
Značka fondu/sbírky:	ŠA
Datace:	(1770) 1848 - 1918
Původce:	Ministerstvo vnitra Vídeň (Ministerium des Innern Wien) 1848-1918
Popis:	Šlechta, nadace, povyšování do šlechtického stavu, povyšování obcí na města, udělování znaků a peče
Metráž (bm):	Nezpracováno: 0 · Zpracováno: 7.87 · Inventarizováno: 7.87
Jazyk:	němčina
Místo vzniku:	Vídeň
Tématické skupiny:	Politická správa po roce 1850

Figure 4: Example of access to information resource applied in Badatelna.eu portal

1.3 Museums and galleries

Museums and art galleries collect unique and rare objects. Their mission is to interpret the unique collections objects rather than simply describe materials for search and retrieval. Their material is unique, therefore local systems for description have historically prevailed and there has been little interest to standardize practices or create systems for shared cataloguing. Therefore, the development and the application of metadata standards and standardized cataloguing do not have a long tradition in museum's environment. Since 1980s, art museum community in the U.S. tried to enforce standardization and to replace the earlier widely varying practices by creation of a common standard. Only in 1995 the first attempt to offer a data structure for museum standardized description Categories for the Description of Works of Art (CDWA) has appeared. CDWA is a standard that provides guidelines for descriptions of works of art – a complex set of 540 categories and subcategories. Since then, CDWA has become a common data structure standard for art museum information and serves as the underlying structure for various home-grown systems. Data content standard Cataloguing Cultural Objects (CCO) based on CDWA appeared in 2005 and since then it has been widely used.

1.3.1 Museums in Czechia

The situation in museums and galleries in Czechia was quite different. Unlike libraries that collect information sources of mass production (books), museums tend to collect unique objects. They create metadata/descriptions of high quality both of unique objects, or their sets (collections) as well, but they apply their own rules valid only for one institution, one collection, one system. Specific content and data structure standards were not developed and published, therefore their options to share metadata/descriptions/catalogue cards were very limited. Moreover, each scientific discipline has its own methodological approach and uses its own terminology in describing and accessing their collections. Museums use highly specialized terminology. A wide range of home-grown indexes and vocabularies are used by the museums and galleries in Czechia from very specialized to very general, but often their terminology is limited to local (one collection, one institution, one system) practice. Common standardized access points do not exist.

System Demus

Demus is a collections management system developed by the Moravian Museum Computer Science Department called CITeM. The advantage of system Demus is well structured data compliant with CIDOC CRM Information Categories (1995). The main drawback of the system is application of outdated database MS Access 97, and use of home-grown, not broadly accepted vocabularies and codes.

System Bach ProMuseum

The Bach ProMuseum system is a system for documenting the collections developed by Bach Systems Ltd. The advantage of the system is the application of indexes, full-text search; the system is easy to use. The main drawback of the system is application of database MS Access, and use of homegrown, not standardized vocabularies and codes.

Central Registry of Museum-type Collections (CES)

CES online database is maintained by the Ministry of Culture of the Czech Republic. It represents an information system of museums' collections accessible to the general public, which offers:

1 A complete overview of museums and galleries founded by the government or by the regional or local municipalities and a list of their collections, briefly characterised and not published anywhere else.
2 Information about the specialised areas in which museums and galleries do their collecting work.

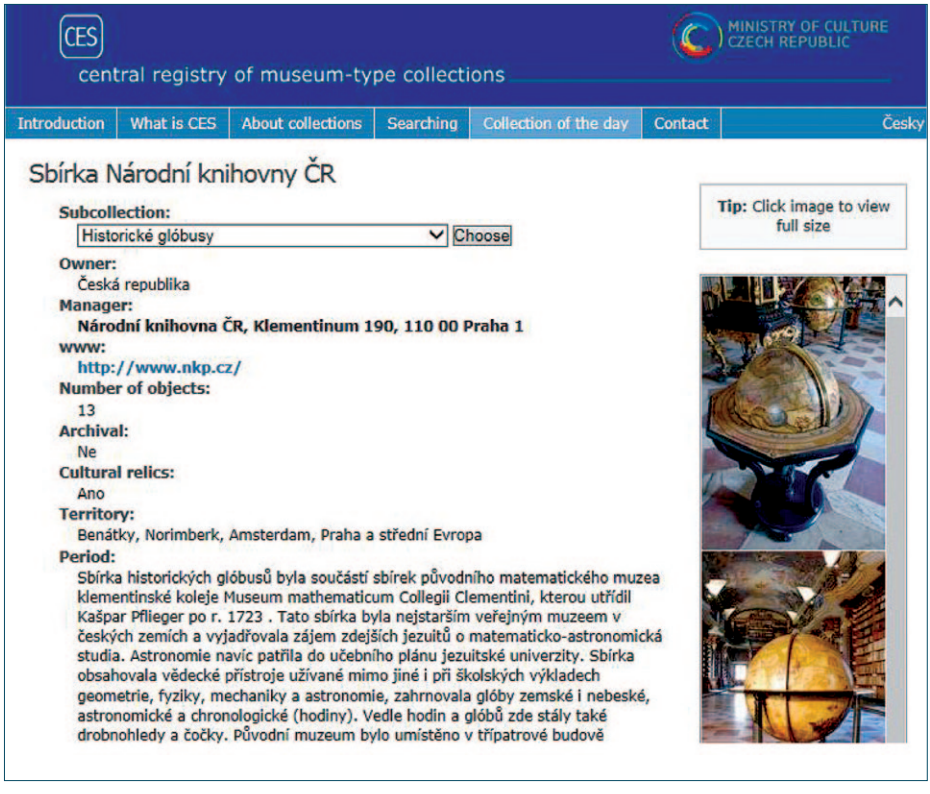


Figure 5: Example of access to information resource applied in CES online database

The database is searchable according:

- the name of the collection
- the owner of the collection
- the manager of the collection
- the subcollections, it means separately recorded parts constituting the collections; each of these parts is focused on a specific area
- characteristics of each subcollection, containing:
- the territory of the subcollection
- the period of the subcollection
- the brief history of the subcollection
- the types of objects and materials
- indication whether the subcollection includes cultural relics or archival documents
- free keywords

Museum Authorities project

Museum Authorities project represents the first attempt to introduce the application of controlled access points – National authority files – in museum environment[1].

The main goals of the project are:

- To establish the operational model for cooperation on the national authority database,
- To test possibilities for harmonization of local thesauri and dictionaries with authority files,
- To create and keep service capacity for maintenance of separate national authority files used by museums.

The cooperation model mainly focused on personal authorities had been designed and implemented in cooperation with the Cosmotron company. Main attributes of the model are:

- Virtual extension of authority records by further information fields, e.g. a broader biography,
- Creating relation from all identifiable entities (places, corporate bodies, persons, terms) used in the particular personal authority record to their respective authority records,
- describing relations among related data (dates, places, corporate bodies, persons, circumstances) through so-called event construction conformable to the conceptual model CIDOC CRM (ISO 21127:2006),
- Relations to supplemental information (especially picture) through URL. (Andrejčíková, 2011, p. 89–90).

[1] Full title of the project: "National Authorities in the Environment of Museums and Galleries – Interoperability with the National Library of the Czech Republic".

Register of Fine Art collections

Register represents the first online union catalogue of museum and gallery objects, it contains 169,177 objects from 18 galleries involved; it is based on National Authorities: personal names data from museum authority files have been used mainly for the unification of different name forms of authors.

Registr sbírek výtvarného umění - detail záznamu
Elektronický katalog sbírek výtvarného umění (pracovní verze)

Řazení: Autor vzestupně
Tabulka: Sbírky << Předchozí záznam Další záznam >>

☐ *Označit tento záznam*

Správce	5NH
Název instituce	Horácká galerie v Novém Městě na Moravě
Skupina	obrazy
Inv. číslo	O 832
Předmět	
Autor	
Titul	Portrét šlechtičny
Podtitul	
Datace	(30. až 40. léta 19. století)
Technika, materiál	olej, plátno
Rozměry	58 x 43,5 cm
Signatura	neznačeno
Rok získání	1992
Námět	portrét

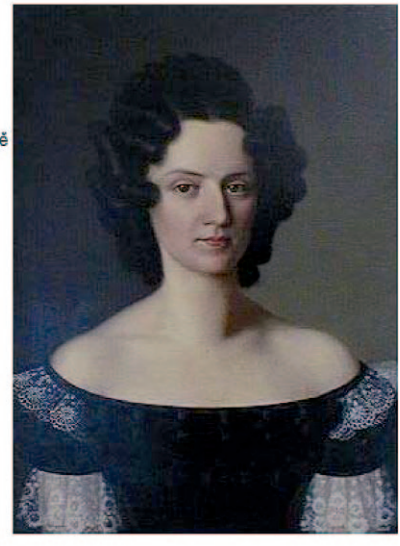


Figure 6: Example of Register of Fine Art collections record

The weakness of all mentioned online databases (PeVa, Badatelna, CES and Register) is that the description of individual collections is simple and that it does not provide a structured description of the individual collection items. At present time, all listed databases are being updated to meet the current needs of users.

2 Information Technologies Context

Since 1980s, resource description and access of cultural heritage information objects/resources relies on the use of information and communication technologies even more than before. The information is recognized as an important resource, and all memory institutions communities must cope with the demands of both developing information society, and emerging new technologies.

Two important standards have been developed, that have changed completely the information technology landscape: a programming language Structured Query Language (SQL) as a special-purpose programming language designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS) (SQL, 2015), and a markup language Standard Generalized Markup Language (SGML) with its successor eXtensible Markup Language (XML) emerged as a subset of SGML in 1998.

Database technology offers representation of data in “regular number of fields; order of fields generally not significant; each field restricted to data; separate fields in which data is entered; interrelated fields have a fixed or shallow hierarchy; and the data in each field is controlled with respect to form and structure. Database technologies excel not only in separating various data fields, but also in reliably the managing the interrelating of the fields allowing users to bring them together in various configurations that serve as a wide variety of perspectives and uses. This approach to data representation is frequently referred to as ‘data-centric’”. (Gueguen, 2013, p. 572)

Markup languages and technologies were intended to “model traditional documents (texts such as essays [...] books) [...] all of these types of data objects share many of the following features: irregular number of components (chapters or paragraphs), sequence or order is important [...] semiregular structure and unbounded hierarchy; arbitrary intermixing of data with markup; and arbitrary number of interrelations within and among documents This approach to data has commonly been called ‘document-centric’”. (Gueguen, 2013, p. 572)

2.1 Application of database technologies in memory institutions

Libraries, as making available non unique information objects (books, journal etc.) to the general public have welcomed the development of integrated library systems (ILS). An ILS contains a relational database, a software to interact with that database, and two graphical user interfaces: for staff and for audience. Most ILSes offer separate functional modules (discrete programmes of software functions) for acquisitions, cataloguing, circulation, and the OPAC (public interface for users). Each module is integrated with a unified interface.

Archives, museums and art galleries, as other memory institutions at that time, remained closed to the general public and mainly dealt with the description of their individual funds (unique information objects). Sophisticated integrated systems for archives, museums and art galleries were not developed and applied.

2.2 Advent of the Internet

In 1980s, Internet as the system of interconnected computer networks (network of networks) was released. In 1990, the Internet represents a global networking infrastructure, in which any computer can communicate with any other computer as long as they are both connected to the Internet. Communication over Internet is enabled by set of rules called as protocols, e.g. TCP/IP (Transmission Control Protocol/Internet Protocol) and HTTP Hypertext Transfer Protocol. The Internet offers a lot of services from which the memory institutions are using mostly the hypertext technology and World Wide Web application.

2.2.1 Hypertext technology

Hypertext is a text with references called hyperlinks to other text. The most important aspect of hypertext is to create links within and among information resources. Hypertext is a powerful and important tool for organization of bibliographic universe. Most important elements like access points in metadata records can be hyperlinked to additional resources, e.g. authority records, bibliographic or factographic databases to provide users with additional information. The current online library catalogues OPACs, bibliographic databases, virtual information systems, portals created by memory institutions benefit from the hypertext technology.

2.2.2 World Wide Web

The World Wide Web (Web) is called Web of documents. It is an application of Internet which creates a virtual network of documents – web pages connected by hyperlink technology. To transmit, the Web uses the HTTP protocol; to access data, it uses browsers, such as Internet Explorer or Firefox. “The model behind the original Web could be summarized as a way to publish documents represented in a standard way (HTML), containing links to other documents and accessible through the Internet using standard protocols (TCP/IP and HTTP). The result could be seen as a worldwide, distributed file system of interconnected documents that humans can read, exchange and discuss” (Introduction to the Semantic Web, 2015).

2.2.3 Semantic Web

The Semantic Web is an extension of the traditional Web of documents. It is called Web of data because it provides an easier way to find, share, reuse and combine information. It enables piece of data to be understood by computer and to be linked from a source to any other source. It means that the Semantic Web connects facts, not documents (web pages only). “The fundamental difference between Semantic Web technologies and other technologies related to data (such as relational databases or the World Wide Web itself) is that the Semantic Web is concerned with the meaning and not the structure of data” (Introduction to the Semantic Web, 2015). It means that information on Semantic Web is understandable for both humans and machines (not only for humans).

From a technical point of view, the Semantic Web consists primarily of following technical standards:

- **RDF (Resource Description Framework)**: The data modelling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF. RDF offered a very different data representation model: graph technologies. “Graph technologies introduce data representation as statements, typically characterized as subject-predicate-object, with each statement called a ‘triple’. While XML supports a specific form of graphs, the hierarchy (or tree) triples enable unbounded representation of networks of interconnected data objects as well as real world objects (represented by data). Given that the real world within which we live and work may be understood as a vast, dynamically interrelated network of people and objects situated in space and time, graph technologies offer new and more expressive forms of representation” (Gueguen, 2013, p. 572).
- **SPARQL (SPARQL Protocol and RDF Query Language)**: The query language of the Semantic Web. It is specifically designed to query data across various systems.
- **OWL (Web Ontology Language)**: The schema language, or knowledge representation language, of the Semantic Web. OWL enables to define concepts in a way that they may/can be reused as much and as often as possible.
- **SKOS (Simple Knowledge Organization System)**: The data model which provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive language for developing and sharing new knowledge organization systems. It may be used on its own, or in combination with formal knowledge representation languages such as OWL. It represents a common data model for sharing and linking knowledge organization systems via the Web. (W3C, 2009)

2.2.4 Linked Data

Linked Data represents a set of best practices for publishing and connecting structured data on the Web. With linked data, related data (piece of data/information), not only related documents, can be found. In the traditional Web of documents (hypertext Web), the relationships between linked documents are implicit, are not machine-readable. Linked Data refers to data published on the Web with explicitly defined meaning according to Linked Data principles defined by Berners-Lee (2006):

1 Use URIs as names for things.
2 Use HTTP URIs so that people can look up those names.
3 When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4 Include links to other URIs, so that they can discover more things.

Linked Data is the de-facto new standard for data publication and interoperability on the Web. The Linked Data paradigm is well positioned to play a strategic role as a pillar of the whole new Internet infrastructure. URIs could be used to uniquely identify real world objects, RDF could be used as a standard way to syndicate data about them and, most importantly, the interlinking could be used to represent objects interactions and relationships with other surrounding objects.

With advent of the new ICT, the memory institutions need to change the traditional procedures, standards, rules and practices. The cultural heritage communities that have long worked largely in isolation from one another, each pursuing its own descriptive practices and developing systems based on these practices, are now increasingly using semantic technologies to explore interconnecting disparate description systems and realize integrated access to cultural heritage. Traditional standards and tools used in memory institutions were developed as organizational tools of physical items (books, objects) in 19th-20th centuries and were designed for the use and consumption of humans. We are sure that the standards and the procedures must be changed, but we do not know how: “Standards that have been in use for decades have come under increasing pressure to either adapt to new circumstances or to give way entirely to different standards. While it is clear that change is happening, what is less clear is where that change is taking us. If e.g. MARC format no longer serves us, then what standards will serve? How can we adapt to fundamental differences in how our data is used without rendering decades of legacy data completely worthless? We stand in a moment of uncomfortable chaos. We must forge a new path, but where that path might lead, or even what it looks like, is still unclear” (Schilling, 2015).

2.3 Metadata standards used by memory institutions

For decades, libraries in Czechia and all memory institutions abroad have been working toward developing data standards creating descriptions of and retrieving information about cultural heritage objects.

Data standards facilitate database creation, promote the recording of information consistently. They may differ in complexity and granularity of fields, they may provide a minimum of agreed elements that facilitate record sharing. Metadata standards facilitate interoperability and legacy resources integration. They support information retrieval.

Some data element sets become formats/schemes by adding rules such as repeatability, controlled vocabularies used, etc.

There are three main types of metadata used in memory institutions in current online environment:

- Descriptive metadata which describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- Structural metadata – provides information how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative metadata – provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; two that are sometimes listed as separate metadata types are:
- rights management metadata, which deals with intellectual property rights,
- and preservation metadata, which contains information needed to archive and preserve a resource.

In INTERMI project, we paid great attention to the descriptive metadata standards used in memory institutions.

Digital cultural content from memory institutions published currently on the Web is of high quality, but very heterogeneous. The processing of information resources in libraries, archives, museums and art galleries was quite different. The domain and application-specific standards and practices have been developed and applied. Because of this diversity – difference in collection type, curatorial approach, subject discipline, granularity and level of detail of description, descriptive data structure, and data content values, – no single descriptive metadata standard has been created that meets the needs of all libraries, museums, and archives. (Gill, 2004)

Heterogeneity of standards is influenced by the heterogeneity of resources to be described. Traditional printed information resources are self-describing: the title, author, place and date of publication can be identified easily. On the contrary, the cataloguing of a unique, museum-type information object requires different approaches, different and more complex standards. The museum-type objects are not self-describing; they do not have a title page that indicates the title, name of author, etc. The information must be created/supplied by the cataloguer. The titles supplied by cataloguers are not stable, they are subject to frequent changes, therefore they are entered in “bibliographic” record, not in authority record only. Therefore, it is almost impossible to create one set of rules applicable to the description of all types of information resources.

Descriptive Metadata standards are divided into:

- Data content standards – provide guidelines for creating metadata.
- Data value standards – represent established lists of terms used as data elements to ensure consistency.
- Data structure standards – metadata schemes define the structure and the meaning (semantics) of elements.
- Data exchange/syntax standards – represent markup languages that provide a standardized way to structure and express metadata schemes for machine processing.

Table 1: Example of Register of Fine Art collections record

	Libraries	Museums	Archives
data content standards	AACR RDA	CCO in Czechia: local, home-grown rules	ISAD (G) ISSAR (CPF) DACS in Czechia: local, home-grown rules Basic Rules for Archival Description (1959) Basic Rules for Archival Description (2013)
data structure	MARC	CDWA	EAD
data format	MARC/XML	XML	XML In Czechia: Proprietary XML
data exchange	OAI Z39.50 SRU/SRV	OAI	OAI
data value standard	Library of Congress Name Authority File (LCNAF) Library of Congress Subject Headings (LCSH) Czech National Name Authorities Czech National Subject Authorities (CZENAS)	The Art & Architecture Thesaurus (AAT), The Union List of Artist Names (ULAN) The Getty Thesaurus of Geographic Names (TGN) Cultural Objects Name Authority Online (CONA)	

2.3.1 Data Content standards

Anglo-American cataloguing rules (AACR)

The rules were developed in 1967, the 2nd Edition Revised was published in 1978 to handle self-describing published textual resources. They were created for defining access points for linear presentation in printed book catalogues or in alphabetically arranged card and first generation of electronic catalogues. Generally speaking, they represent descriptive rules for traditional textual sources (books) in the traditional environment. First non-traditional formats – musical recordings and motion pictures – were integrated as well by entering all significant differences mostly in the notes and physical description areas. Coyle and Hillmann (2007) acknowledge that the situation changed basically with the explosion of digital formats, the Internet and its application World Wide Web. The description and documentation of new emerging digital, non-book formats using the model of book cataloguing became less useful, because the descriptive rules were based on and defined for predictable, stable and named sources of information (title pages, colophons, etc.), with a prescribed order of preference; these rules were not adaptable to resources without title pages or pages, and not suitable for resources that existed in a state of constant change.

Resource Description and Access (RDA)

The successor of the AACR2 rules, Resource Description and Access (RDA) was published in 2010. RDA has been developed to replace the Anglo-American Cataloguing Rules. RDA builds on AACR2 foundations, but it is a new standard for resource description and access, designed for the digital world. RDA has been designed to be compatible with legacy AACR2 records. “Built on foundations established by the Anglo-American Cataloguing Rules (AACR), RDA will provide a comprehensive set of guidelines and instructions on resource description and access covering all types of content and media. The new standard is being developed for use primarily in libraries, but consultations are being undertaken with other communities (archives, museums, publishers, etc.) in an effort to attain an effective level of alignment between RDA and the metadata standards used in those communities.” (Joint Steering Commitee for Development of RDA, 2009, p. 1) RDA is based on two international conceptual models, FRBR and FRAD from which, RDA gets the entities, the identifying attributes for each entity, the relationships, and the user tasks.

In Czechia, RDA rules have been applied since May 2015. As the INTERMI project started in 2011, we had to take into account both rules AACR and RDA as well, and to identify differences between them.

Differences between RDA and AACR2

- In RDA, there is no longer a rule of three applied:
  In AACR2, if there were three names or less on the title page, all should be entered. If there were more than three, only the first one was entered.

In RDA, it is recommended to list all of the authors, along with any additional information that appears with the name. However, it is only required to give the first statement of responsibility.

- Providing access to contributors has changed:
  In AACR2, access to the authors listed in the statement of responsibility is provided. In RDA, it is required to provide an access point for the creator of the work, but all other access points for names are optional.
- Abbreviations are not used any more:
  RDA only uses abbreviations when they appear as such on the work itself (i.e. the title page has “v.” or “Dept.” on it.) Furthermore, Latin abbreviations will no longer be used, unless they appear on the work itself.
- The GMD [general material designation] is provided:
  In AACR2, this information appeared near the title. In RDA, this element is being replaced by three new fields. These fields will supply information on the content, the media, and the carrier.
- Typographical errors and inaccuracies:
  In AACR2, the convention of [sic] or [i.e.] when there was a typographical error in the text, is used. In RDA, the inaccuracy is transcribed as it appears and then a note is added.
- Unique identification of works:
  In AACR2, serials were required to have unique titles, monographs were not. In RDA, every work is required to have a unique authorized access point.
- Other changes:
  There are significant changes in formulating access points for Bible, and for fictional characters.
  RDA has more requirements for minimal bibliographic records than AACR2.

Cataloguing of Cultural Objects: a Guide to Describing Cultural Works and Their Images (CCO)

It represents a manual for describing, documenting, and cataloguing cultural information objects: works and their visual surrogates. It is a data content standard for cultural information objects including paintings, sculpture, prints, photographs, built works, installations, and other visual media, artefacts, archaeological sites, manuscripts. CCO provides guidelines for descriptive metadata for unique items, not bibliographic items nor Web resources. It identifies 116 elements in total, from which 9 are considered essential for the identification of unique, museum-type objects (core elements). It was developed independently from other data content standards, it is the first comprehensive cataloguing standard which did not originate in the library community environment. It is more complex than other cataloguing standards: it provides instruction for display as well as indexing forms and includes chapters on authorities. Other important feature is that it is data format independent. Detailed instructions on the use of this standard are formulated in ten key principles. (Visual Resource Association, 2015)

CCO recommends the use of controlled vocabularies and thesauri such as The Art & Architecture Thesaurus (AAT), The Union List of Artist Names (ULAN) and The Getty Thesaurus of Geographic Names (TGN), as well as others. With CCO, the museum and visual resources communities now have a data content standard specific to the materials they describe. (Elings, Waibel, 2007)

International Standard Archival Description (General) – ISAD (G)

ISAD (G) is a standard which gives guidelines for describing archival materials. It was developed by a Committee of the International Council on Archives with the aim to standardize the archival description at international level. By creating the standard, the Committee respected existing national standards for archival description and best practices whenever possible. It is to be used in conjunction with the existing national standards or as basis for the development of national standards. ISAD (G) promotes the creation of consistent and appropriate descriptions, supporting the retrieval and exchange of information, and the integration of descriptions into a unified information system. It identifies a set of elements, rules which could be applied, and describes what kind of information should be included in the archival description. It contains 26 elements, from which 6 are considered essential for the international exchange of information: reference code, title, creator, dates, extent of the unit of description, and level of description. For multi-level descriptions, ISAD (G) recommends to provide the description from general to more specific and to locate each description within the hierarchy. (Data Structure Standards, 2013)

International Standard Archival Authority Record (Corporate Bodies, Persons and Families)[2]

ISAAR (CPF) is a standard which gives guidance on preparing archival authority records that describe corporate bodies, persons and families associated with the creation and maintenance of archives. It allows for the development of authorised name forms, which can assist users in interpreting and understanding the value of the records created by the person, family or corporate body they are interested in.

Authority records can be shared and linked together more easily if they have been developed in a standardised way. ISAAR (CPF) makes recommendations for the standardised creation of these records and offers a model by which they can be linked to descriptions of records and to other information resources related to the records creator.

Both international standards ISAD (G) and ISAAR (CPF) represent a base for creation of the national standard for archival description, e.g. DACS, and Basic rules for archival processing in Czechia.

[2] Available at: *http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf*

Describing Archives: A Content Standard (DACS) – national application of ISAD (G)

DACS is an output-neutral set of rules for describing archival materials. It was developed because AACR2 did not fully meet the needs for description of archival collections to capture the complexity of archival materials. DACS was officially approved by the Society of American Archivists as a SAA standard in March 2004 and since then it has been used in U. S. archives as a replacement for Archives, Personal Papers, and Manuscripts (APPM). It represents the U. S. implementation of international standards (ISAD(G) and ISAAR(CPF)) for the description of archival materials and their creators. It is used to describe archival materials at any level of specificity, from the collection to the item level. DACS represents a complex standard: it provides an overview of archival description (including the requirements for effective multilevel description), outlines the elements that must be included at different levels of description, and describes how those elements should be implemented. DACS is based on the same principles found in the new Resource Description and Access. In addition, DACS provides specific guidance in describing creators of archival material, constructing archival authority records, and recording forms of names.

Basic Rules of Archival Processing

Basic Rules for Archival processing provide a comprehensive set of rules for describing archival materials; it standardizes and unifies the archival description in order to ensure the archival information to be accessible in both at national and international level as well. It is meant as a replacement of the Basic Rules of Archival Processing of 1959 published in 1960.

It is based on international standards for archival descriptions, e.g. the International Standard for Archival Description ISAD (G), the International Standard Archival Authority Record for Corporate Bodies, Persons and Families ISAAR (CPF) and the International Standard for Describing Institutions with Archival Holdings ISDIAH.

The main goal of the standards is to ensure the objectiveness, precision, and standardization of the archival description applied in Czech archives of all types.

Factors influencing the creation of national rules for archival description in Czechia include:

- the need of application of international standards in Czechia,
- the emergence of new types of archival materials (digital born documents, technical documentation),
- the need to harmonize description of archival content in Czechia,
- the emergence of new, untraditional types of originators to which the traditional criteria of definition are not suited (international companies, multi-level originators),
- the new form of registries and registry systems that are not supported by former Basic Rules 1959),

- the impact of the emergence of new information and communication technologies above all Internet, Web, LOD,
- the user requirements for effective online access to archival content,
- and current/future collaboration among memory institutions in Czechia.

The main goal of the rules is to define the context of archival processing, types of archival records, definition of archival sets, application of provenance and pertinence principles, and definition of a specifically Czech archive feature – "registration units" ("evidenční jednotky")[3]. Further, Basic rules include the levels of archival descriptions, entities dealing with origin of archival materials (originators) and their attributes and relationships, rules for formulation of access points and descriptions of archive and cultural research institutions that keep archival records (museums, libraries, galleries, memorials, public research institutions, and universities). Basic rules contain a set of examples as well.

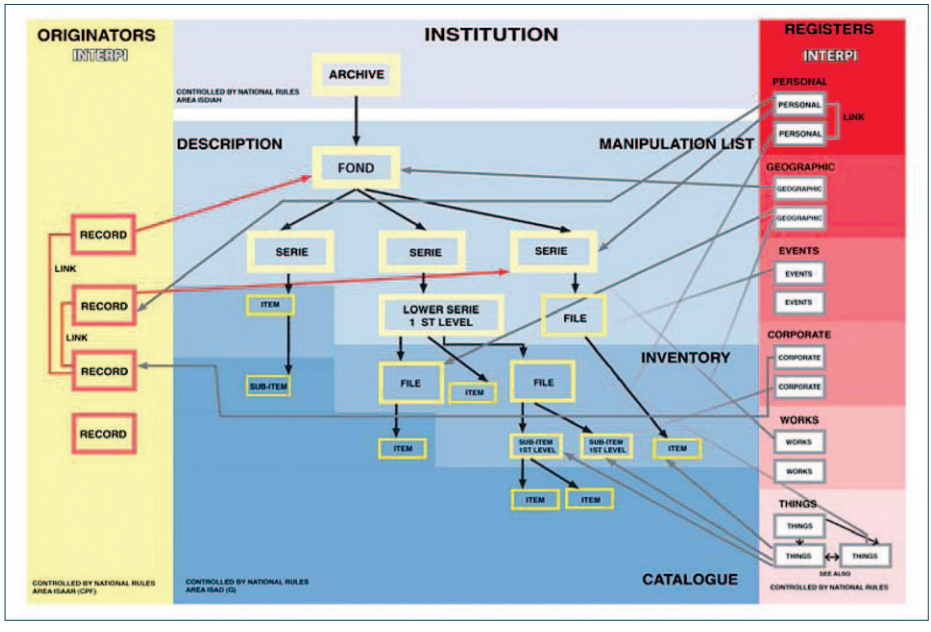


Figure 7: Schema applied in Basic Rules of Archival Processing

2.3.2 Data structure standards – metadata schemes applied in memory institutions

Metadata schemes define the structure and the meaning (semantics) of data elements organized in databases in such a way that it can be used efficiently.

[3] The main objective of Registration unit is to link finding aids to the Registers of the National Archival Heritage (NAH).

MARC format

MARC format was developed and implemented just to share bibliographic information in electronic form. It began in the mid-1960s as an LC project with the aim of making use of bibliographic data in a machine-readable form. The main purpose of the standard has been to enable the cataloguing of books, the exchange of authority files, and the creation of union catalogues. MARC supports eight types of materials:

- books (printed, electronic, manuscript and microform textual materials),
- continuing resources (textual materials that are issued in parts with a recurring pattern of publication),
- computer files (computer software, numeric data, computer-oriented multimedia, online systems or services),
- maps (printed, electronic, manuscript, and microform cartographic materials),
- music (printed, electronic, manuscripts, and microform music, as well as musical sound recordings or other sound recordings),
- visual materials (projected media, non-projected media, two-dimensional graphics, three-dimensional graphics, naturally occurring objects),
- and mixed materials (mixture of forms of materials).

The format was adopted or adapted by numerous libraries around the world as data exchange format and „working format" as well. There are two formats in use today: MARC21 format which is more used and supported – maintaining and controlling agency over Format is the Library of Congress[4], and UNIMARC format which is widely used in Europe and it is supported by Permanent UNIMARC Committee (PUC)[5].

MARC 21 format is widely used in library community. OCLC database comprises 200 million records from 70,000 libraries around the world, Library of Congress comprises 12 million records, National Library of the Czech Republic comprises 2,374,963, Union catalogue (Czech Republic) comprises 6,135,996 records.

A typical MARC record (bibliographic and authority as well) consists of multiple fields that are marked with tags consisting of a three-digit code which describe what kind of data is within this field. Some fields are defined by indicators. Indicators contain values conveying information that interprets or supplements the data found in the field. Indicators are defined independently for each field. The fields can be further divided into subfields that are separated by a delimiter which is complemented by a subfield code indicating what kind of information will follow. Subfield codes identify data elements within a field that require (or might require) separate manipulation and may be any lowercase

[4] For more information: *http://www.loc.gov/marc/*.

[5] For more information: *http://www.ifla.org/publications/unimarc-formats-and-related-documentation*.

alphabetic or numeric character. Subfield codes are defined independently for each field; parallel meanings are preserved whenever possible. (The MARC 21 Formats, 1996)

Application of MARC format in current online environment

The basic problem is that MARC format had not been accepted by other memory communities, so it is not a mainstream format. The MARC format has been found too complex, not flexible enough; its main purpose was to provide shareable bibliographic data for regularly published works. It was not developed for sharing the complex and structured archival and museum descriptions. The MARC formats serve as standards for the representation and communication of bibliographic information between systems on the record level. MARC records involve three elements: the record structure, content designation, and data content. Data elements within a MARC record are identified by codes and conventions (Guenther, 1997). It was assumed that the codes and conventions are able to clearly express the meaning of data and support easy manipulation of those data. But in the case of bibliographic data stored in MARC format, it is quite a difficult task. The MARC record was not created as a set of data elements but as a format for the storage and display of the text of library catalogue records and to be understandable for humans. Coyle (2011) acknowledges that one of the difficult tasks in analysing MARC21 is to separate the record structure from its content, as the two are not distinguished in the format itself. MARC21 records contain some punctuation signs to visually delimit different data elements or in order to concatenate some elements for printing without the need for additional programmatic formatting (ISBD punctuation). The meaning of data is influenced by indicators. It is considered rather a mark-up language for the catalogue record text. Library data is stored in a format which is not compatible with current standards used on the web or the Semantic Web. Research studies shown that the transformation of library data to be more ready for the re-use outside of the library community would be very demanding. (Park, Kipp, 2014)

Bibliographic Framework (BIBFRAME)

BIBFRAME is a new model for bibliographic universum proposed by Library of Congress created with the aim to publish/expose bibliographic data on the semantic Web. The aim of BIBFRAME Initiative is to develop a foundation for the effective representation, exchange and exposition of library data on the Semantic Web. Current library MARC data depend explicitly on the use of lexical strings for identifiers of bibliographic entities (name of author, title...). The new model offers the creation and application of clearly identified entities and relationships between them and the use of machine-friendly identifiers for those entities and relationships with the aim to ensure the machine interpretation of both of them. The aim is to replace MARC format, to accommodate different content models and cataloguing rules, to explore new methods of data entry, and to evaluate current exchange protocols.

There are four high-level classes, or entities, in the BIBFRAME Model:

- BIBFRAME Work.
- BIBFRAME Instance.
- BIBFRAME Authority.
- BIBFRAME Annotation.

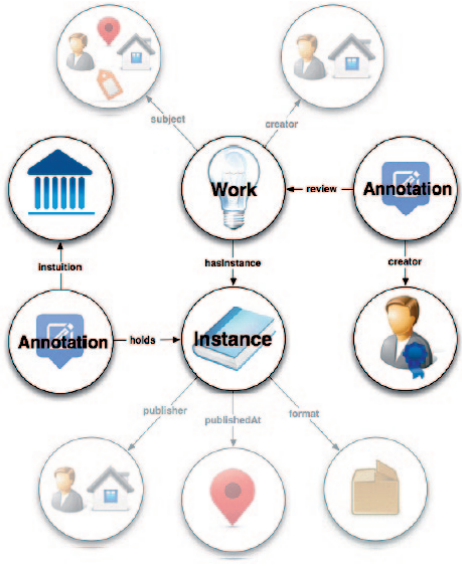


Figure 8: A graphical representation of the BIBFRAME Linked data model (Library of Congress, 2012)

BIBFRAME Work identifies the conceptual essence of a thing; a BIBFRAME Instance reflects the material embodiment of a Work; a BIBFRAME Authority identifies a thing or concept associated with a BIBFRAME Work or Instance; and a BIBFRAME Annotation provides a new way to expand the description of a BIBFRAME Work, Instance, or Authority.

The BIBFRAME Model relies on relationships between resources (Work-to-Work relationships; Work-to-Instance relationships; Work-to-Authority relationships). It manages this by using controlled identifiers for things (people, places, languages, etc.) (Library of Congress, 2012).

Categories for the Description of Works of Art (CDWA)

CDWA are a set of guidelines for best practice in cataloguing and describing works of art, architecture, other material culture, groups and collections of works, and related images. CDWA represent a conceptual framework that may be used for designing databases and accessing information. They include both a conceptual framework of elements and relationships, and cataloguing rules for describing, documenting, and cataloguing cultural works and related images. CDWA contain around 540 categories and subcategories of information for describing art domain. The exhaustive data structure developed originally from the point of view of academic researchers may be customized for different audiences as well.

In academic research community, CDWA is regarded as core representing the minimum information necessary to identify a particular work of art or museum object. Other communities do not require such a large amount of information, they use simpler structures based on CDWA as CDWA Lite, VRA core categories and LIDO.

CDWA Lite

CDWA Lite is an XML schema to describe core records for works of art and material culture based on CDWA and CCO. CDWA Lite records are intended for contribution to union catalogues and other repositories using the Open Archives Initiative (OAI) harvesting protocol. (Getty, 2015)

VRA core categories (VRA)

VRA is a metadata schema based directly on CDWA that focuses on describing visual documents or visual surrogates of art, architecture, and material culture. It is used mostly by librarians of slide collections who need to describe the object as well as slide, digital image, etc.

It was developed in 1996, current version was released in 2007 and it is expressed as XML schema. It recognizes three main entities: collection, work, and image, and relationships between them. Collection in VRA Core is an "aggregate of work or image records" (Rose, 2006).

A work is a physical entity that exists, has existed at some time in the past, or that could exist in the future. It might be an artistic creation such as a painting or a sculpture, a performance, composition, or literary work, a building or other construction in the built environment, an object of material culture. Works may be simple or complex. Works may have component parts that are catalogued as works themselves but related to the larger work in a whole/part or hierarchical fashion.

An image is a visual representation of a work. It can exist in photomechanical, photographic and digital formats. In a typical visual resources collection, an image is a reproduction of the work that is owned by the cataloguing institution and is typically a slide, photograph, or digital file. A visual resources collection may own several images of a given work.

Encoded Archival Description (EAD)

EAD is a non-proprietary de facto standard for the encoding of finding aids for use in a networked (online) environment. It was developed by the library of the University of California at Berkeley to meet all needs for archival description and to include more information than was provided by MARC records. The special needs include ability to:

- maintain the hierarchical relationships between the different levels of description,
- provide descriptive information independent of these inherited hierarchical levels,
- navigate within such a hierarchy of structured information,
- and support for element-specific indexing and navigation.

EAD allows the standardization of collection information in finding aids within and across repositories and can be encoded in XML. The EAD elements are specified in the EAD Document Type Definition (DTD). The tag set consists of 146 elements and is used both to describe a collection as a whole, and for the encoding of a detailed multi-level inventory of that collection.

2.3.3 Conceptual Data Models developed by memory institutions communities

It is well known that information managed and published by memory institutions (libraries, museums and archives) are heterogeneous in structure and content, although there are significant conceptual overlaps. In 1990s, memory institutions communities decided to try to overcome this diversity by applying conceptual modelling techniques in defining data models which could better represent its data. International institutions (IFLA, ICOM)[6] as representatives of libraries and museums communities initiated the development of conceptual models at the international level.

The "Library community" conceptual data models were based on users' requirements and were designed to provide a high level conceptual view of the domain covered by bibliographic databases. Along with a conceptual model for museum community CIDOC CRM represent abstract, general models. As the implementation methodology was not created and developed (FRBR, CIDOC CRM), they are not "actionable", they are not ready for use in databases and programs. The documentation does not provide data creation rules, the definitions of entities, attributes and relationships between them are not specific enough and in a way that could be transformed easily into specific rules or programmed into algorithms (Coyle, 2015). They can serve as:

- tools for evaluation of existing rules for description, formats, data models, and to improve them,
- tools for creation and development of mediation tools between heterogeneous databases (i.e., databases, that do not have the same format, or not the same data model, or are not based on the same rules for description),
- and ontologies with the aim to contribute to the development of the "Semantic Web".

The FRBR family of conceptual model models (FRBR, FRAD, FRSAD) $FRBR_{OO}$

Conceptual modelling is one of the formal techniques for representing the principal concepts and relationships between and among them for a given knowledge domain. In 1990s, most popular method for database technologies was the entity-relationship model (ER). This method was used in the development of "library" conceptual models: FRBR, FRAD, FRSAD.

[6] ICOM – International Committee for Documentation

Functional Requirements for Bibliographic Records (FRBR)

Functional requirements for bibliographic records (FRBR) represent a conceptual model of the bibliographic universe developed by the International Federation of Library Associations and Institutions (IFLA). It was published in 1998. The FRBR model was developed by the library community to better serve library users in discovering information. The FRBR model is based on the entity-relationship analysis technique common in database design. It is highly theoretical. "It is not a data model, it is not a metadata scheme, it is not a system design, but rather an abstract model of all the things that libraries, museums, and archives collect for our users" (Tillet, 2004, p. 197). It is intended to be independent of any cataloguing code or implementation (system-neutral, and code-neutral). It defines the key objects in a bibliographic domain: entities, attributes, and relationships at as high level as possible, but the important information as definitions of values for some attributes or definitions of some relationships are not included.

Current experience shows that the model is more useful for library users who are looking for works with many expressions and manifestations (books in library catalogues, which represent products of mass production); it is not as helpful for works with one expression and very few manifestations (manuscripts) or single-objects entities like artworks and other museums information objects which are related to many other information objects.

Functional Requirements for Authority Data (FRAD)

FRAD was published in 2009 by the International Federation of Library Associations and Institutions (IFLA). It is a highly theoretical, entity-relationship (E-R) model for authority data. The main goal of FRAD is "to provide a framework for the analysis of functional requirements for the kind of authority data that is required to support authority control and for the international sharing of authority data." The FRAD model identifies and examines four tasks employed by users (cataloguers, librarians, patrons) of authority data: find, identify, contextualize, and justify. There are sixteen FRAD entities on which authority data are focused. Included are the ten entities outlined in the FRBR model (i.e. person, corporate body, work, expression, manifestation, item, concept, object, event, place) plus six new entities (i.e. family, name, identifier, controlled access point, rules, and agency). Each entity is associated with its own prescribed set of attributes. "In FRAD, authority data collocates works by a person, family, or corporate body, or the various editions of a title, by providing the controlled access points and variant forms of a name." (Jin, 2012)

Functional Requirements for Subject Authority Data (FRSAD)

FRSAD is a conceptual entity-relationship model developed by International Federation of Library Associations and Institutions (IFLA). This model focuses on the relationships between a Work, its subjects, the way these subjects are named, and the information contained in indexing schemes about both the concepts and the appellations that refer to them. Model defines four specific user

tasks of subject authority data:

- Find one or more subjects and/or their appellations that correspond(s) to the user's stated criteria, using attributes and relationships.
- Identify a subject and/or its appellation based on its attributes or relationships (i.e., to distinguish between two or more subjects or appellations with similar characteristics and to confirm that the appropriate subject or appellation has been found).
- Select a subject and/or its appellation appropriate to the user's needs (i.e., to choose or reject based on the user's requirements and needs).
- Explore relationships between subjects and/or their appellations (e.g., to explore relationships in order to understand the structure of a subject domain and its terminology). (IFLA, 2010, p. 34)

FRSAD model contains only two entities, nomen and thema. Anything (e.g. person, corporate body, place, event, concept, etc.) that can be the subject of the FRBR entity work is known as a thema. Nomen is "any sign or sequence of signs (alphanumeric characters, symbols, sound, etc.) by which a thema is known, referred to, or addressed as" (IFLA, 2010, p. 15). Each entity of FRSAD's model is prescribed an individual set of attributes.

Figure 9: FRSAD conceptual model

Many-to-many relationships exist between the entities of work and thema as well as those between the entities of thema and nomen. "This means that any work may have several themas and any thema may be a subject of several works" (Zumer, Zeng, Salaba, 2012, p. 46). Several nomens may apply to a thema and viceversa: more than one thema may share the same nomen. Qualifiers are used to disambiguate identically spelled nomens from one another that are included in a controlled vocabulary so as to provide clarification for each nomen's specific meaning.

All three conceptual E-R models, FRBR, FRAD and FRSAD, provide the foundational structure for RDA (Resource Description and Access) cataloguing standards. FRBR and FRAD assist in the identification of relationships between a work and its creators as well as a work's various editions and formats. "The contribution of FRSAD to the bibliographic universe and beyond is its applicability for both subject and non-subject authority data that are used for expressing either aboutness or ofness of intellectual and artistic works." (Zumer, Zeng, Salaba, 2012, p. 104)

Object-Oriented formulation of FRBR – FRBR$_{OO}$

The FRBR model was originally designed as an entity-relationship model. It was later – in collaboration with CIDOC CRM community – reformulated as an object-oriented model. As a consequence, there are currently two distinct

versions of the FRBR model: $FRBR_{ER}$ and $FRBR_{OO}$. The first version of the resulting model, $FRBR_{OO}$ was published in 2009, the second one in 2013. It uses the formalism of CIDOC CRM, that adds to the original FRBR the dynamic aspects of the model. "$FRBR_{OO}$ was developed with Semantic Web technologies in mind, and lends itself well to the Linked Data environment" (Le Boeuf, 2012). The document itself is very complex and demanding to understand because of semantic ambiguity and duality in defining some concepts.

Differences in formalism between $FRBR_{OO}$ and FRBR, FRAD, FRSAD

User tasks

FRBR, FRAD and FRSAD define user tasks, entities, attributes and relationships between entities. $FRBR_{OO}$ does not explicitly mention any user tasks, because the issue of user tasks does not form an integral part of any conceptual model. User tasks should be formulated in a cataloguing code, not in a conceptual model. Le Boeuf (2015) acknowledges that all FRBR family conceptual models should declare only such information elements that are likely to enable users to:

- find precisely what they were looking for,
- find what they were not looking for but it is related to it in some way (i.e., explore),
- concept and, to a lesser degree, trust the information they have found.

Naming conventions

$FRBR_{OO}$ uses different naming convention as FRBR, FRAD, and FRSAD do.

The term "entity" was replaced by the term of "class". The identifier for a class consists of the letter "F" followed by a number, e.g. "F1" and a label in a natural language e.g. "Work".

"Scope note" represents a very important information tool in the model: it should accompany the declaration of every single class (and property as well) with the aim to explain how the class and property should be used, e.g. "F11 – Corporate Body".

F11 Corporate Body

Subclass of:	E74 Group
Superclass of:	
Scope note:	This class comprises organisations and groups of two or more people and/or organisations acting as a unit. To be considered an F11 Corporate Body a gathering of people needs to bear a name and exhibit organisational characteristics sufficient to allow the body as a whole to participate in the creation, modification or production of an E73 Information Object. Groups such as conferences, congresses, expeditions, exhibitions, festivals, fairs, etc. are modelled as F11 Corporate Bodies when they are named and can take collective action, such as approving a report or publishing their proceedings.
Examples:	The International Machaut Society The British Library

Figure 10: Example of class descriptions: Corporate Body

Transforming FRBR/FRAD/FRSAD attributes into properties

The CIDOC CRM model declares no "attributes" at all (except implicitly in its "scope notes" for classes), but regards any information element as a "property" (or "relationship") between two classes. The semantics extracted from $FRBR_{ER}$ attributes are therefore rendered in $FRBR_{OO}$ as properties, according to same principles as the CIDOC CRM model.

The identifier for a property consists of the letter "R" followed by a number, e.g. "R7" and a label in a natural language e.g. "is example (has example)". Some properties are identified by "CLP", which stands for "Class Property".

Many notions that were declared as entity attributes in FRBR, FRAD, FRSAD models are modelled now as classes in $FRBR_{OO}$, because it appeared that they were more complex than initially assumed e.g. Date.

Introduction of temporal entities, events and time processes

Temporal entities play a central role in CIDOC CRM model, as they are only means to relate objects to time-span, locations, and agents. Temporal entities were introduced into $FRBR_{OO}$ by declaring some of the classes of $FRBR_{OO}$ e.g. "F28 Expression Creation" as subclasses of classes from CIDOC CRM: "E65 Creation", "E12 Production".

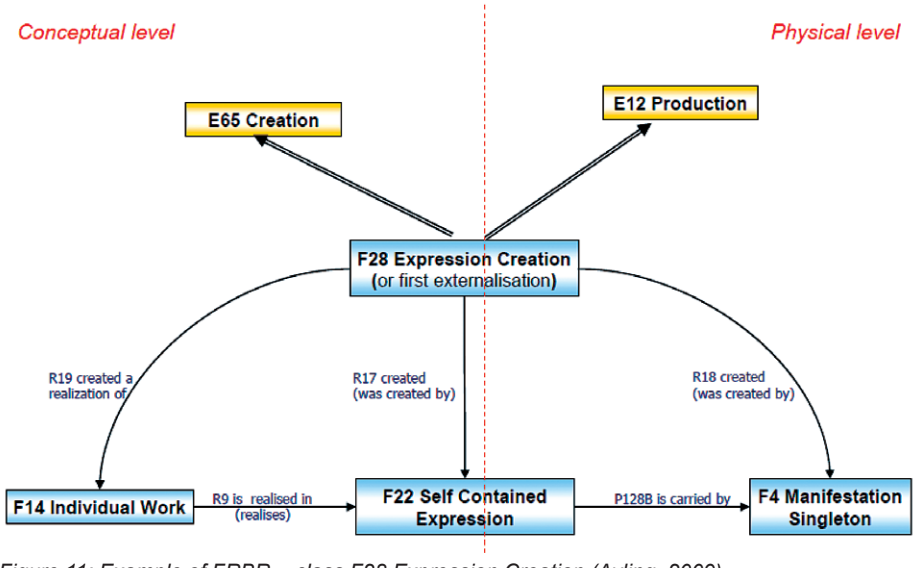


Figure 11: Example of $FRBR_{OO}$ class F28 Expression Creation (Ayling, 2009)

Refinement of group 1 entities

$FRBR_{OO}$ retains the notion of "Work" (in $FRBR_{ER}$ key entity of group 1) as a superclass for the various possible ways of interpreting the $FRBR_{ER}$ definitions; it introduces new classes F14 Individual Work; F15 Complex Work; F16 Container Work; E19 Publication Work.

F1	-	-	-	-	-	-	Work		
F14	-	-	-	-	-	-	-	Individual Work	
F17	-	-	-	-	-	-	-	-	Aggregation Work
F15	-	-	-	-	-	-	-	Complex Work	
F18	-	-	-	-	-	-	-	-	Serial Work
F16	-	-	-	-	-	-	-	Container Work	
F19	-	-	-	-	-	-	-	-	Publication Work
F18	-	-	-	-	-	-	-	-	- *Serial Work*
F20	-	-	-	-	-	-	-	-	Performance Work
F21	-	-	-	-	-	-	-	-	Recording Work
F17	-	-	-	-	-	-	-	-	*Aggregation Work*

Figure 12: Example of $FRBR_{OO}$ class hierarchy: Work

Enrichment of CIDOC CRM from $FRBR_{OO}$

The analysis provided for bibliographic processes in $FRBR_{OO}$ indicates the way for the introduction of refinements into CIDOC CRM, so that the museum community's model has now the opportunity to "model of mass production phenomena as they apply to certain categories of objects found in museum collections (such as the printing of engravings), or the relationship between the creation of content (which is immaterial) and its physical carrier. Further, it introduced a basic model of intellectual conception and derivation applicable to all art forms. This required that the concept of work, among others, be integrated into CIDOC CRM" (Riva, Doerr, Zumer, 2008).

2.3.4 CIDOC CRM

The CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation (ICOM, 2014). In 2006, CIDOC CRM was accepted as the ISO 21127 standard.

CIDOC CRM is an object-oriented ontology which formalizes the semantic concepts used in museum, library and archive documentation applying object-oriented data modelling techniques. It enables an explicit definition of cultural event and object attributes, as well as of their relationships. It is high-level ontology, too abstract to be used as an information model. On the contrary, "it is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives" (ICOM, 2014).

CIDOC CRM is event-centric ontology; it is based on the assumption that "event-centric documentation provides a more accurate view of the past or current life history of a cultural object" (Doerr, Kritsotaki, 2006).

The structure of the CIDOC CRM is based on a class hierarchy of 94 named classes, interlinked by 168 named properties. As in other object-oriented systems, the classes in the hierarchy inherit properties from their parents; in this case, they represent superclasses.

Conceptual model for archival description

Conceptual model for archival description has not been developed yet.

In late 2012, the International Council on Archives (ICA) appointed the Experts Group on Archival Description (EGAD) to develop a conceptual model for archival description that would integrate the four existing descriptive standards: ISAD(G), ISAAR (CPF) ISDF and ISDIAH. This model should ensure consistent, relevant and explicit descriptions, facilitating the retrieval and exchange of information about archival materials, and making it possible to integrate descriptions from different repositories into unified information.

It should take full advantage of opportunities presented by currently emerging communication technologies including the opportunities to work cooperatively within and outside the archival community in a shared information landscape. (Gueguen, 2013, p. 566)

2.4 Conclusion – what we have learned

1 Collaboration among memory institutions is a key aspect of further development

Cultural heritage landscape is very rich on information but highly diverse. Each community uses its methodology, proprietary standards and data structures. The situation in Czechia is more complex. As it was shown, some institutions (libraries) are able to use sophisticated systems, structures and standards, but their data is understandable only or mostly for humans. The libraries are not able to expose their data on the Web to be shareable for machines: so their data must undergo demanding transformation.

Other communities do not use common standards or rules, as these have not been developed yet or were developed recently. Some of the standards have been completed in close collaboration with INTERMI project partners.

2 To identify and specify key elements of bibliographic, archival and museums' universe using conceptual modelling techniques

The second and important task is to specify key elements that are used in memory institutions when describing information resources. To do this, we use conceptual modelling technics and conceptual models designed in the 1990s, and their gradual harmonizations (FRBR family models, CIDOC CRM and $FRBR_{OO}$). Here, however, we lack a conceptual model for archival universe, which is still evolving. Based on these analyses, key categories (classes) and common basic relationships between them were identified and defined.

3 Object oriented principle and event-centric approach should be introduced into the INTERRMI conceptual model

When defining proprietary INTERMI conceptual model we respected two important principles: object oriented principle and event-centric approach (influenced by CIDOC CRM). It was necessary to create a proprietary model, so that the needs of all memory institutions would be met; in addition, the existing conceptual models are rather theoretical and abstract, not containing recommendations and rules for the creation of an information system.

4 As data content standards, RDA rules will be accepted by all memory communities

Museum and archive communities in Czechia do not have their own data content standards developed. After a detailed analysis they accepted the solution offered them by INTERMI team to apply RDA rules which provide a comprehensive set of guidelines and instructions on resource description covering all types of contents. However, for specific areas of entities description, e.g. description of entity appellation, description of dates or definition of rules for mapping of controlled vocabularies used in memory institutions, specific instructions will be formulated. This way, all needs of memory institutions communities will be respected.

5 National authorities will integrate all needed information and will be used in INTERMI project

Museum and archive communities in Czechia do not have their own data value standard developed as well. They agree to apply both National name and Subject authority files by describing their information objects; however, assuming that the national authorities will be modified and extended to include the necessary information to meet all memory institutions communities needs.

6 MARC format will not be used as common INTERMI format. INTERMI format must be compatible with current standards used on the web or the Semantic Web

Regarding data structure format, MARC bibliographic format was found to be old-fashioned, extremely complex and inadequate to the archival and museums objects description.

3 INTERMI conceptual model

Introduction and context

Libraries, archives, galleries, and museums as memory institutions collect, organize and make available information objects/resources of national cultural, industrial, scientific, and natural heritage domain. Cultural heritage domain includes documentary and archival heritage as well.

Now, in the period of the mass digitisation, the main goal of memory institutions activities is to support effective selection, aggregation and easy use/reuse of the digital content.

In INTERMI project, this common objective depends on the specification of user community needs, on the quality of metadata related to the information objects (unambiguous identification of objects on the metadata level included), on application of object-oriented programming principles and event-centric approach, on application of semantic interoperability technologies, it means on conceptual model and data structure applied, on formal representation tools suitable for Semantic Web environment (e.g. XML, RDF, OWL), and on application of Linked Open Data principles.

Specification of user's community needs and their expectations from INTERMI project

Users of memory institutions in Czechia (general public, cataloguers, indexers, information specialists, curators, archivists, and professionals from all subject domains) support the creation and development of the infrastructure for cultural heritage database; they prefer to make cultural heritage content available on the Web in both human readable and machine understandable form, it means to share cultural heritage content according to Semantic Web and Linked Open Data principles. They are aware that it supposes development and application of standards, rules and metadata schemas of high quality and to be appropriate for the Semantic Web. They need more information about entities in current authority records and they agree that it is necessary to move from authorities to entities of value on the web, from terms to concepts.

Information objects and their identification on metadata level

In INTERMI project, the information objects represent the real-world entities, e.g. persons, institutions, three-dimensional objects (i.e. artistic and technical objects, objects of inanimate nature), activities, events and performances, artistic and other achievements, places, etc. which represent topics of their collection items and are subject to their activities. At the time being, the identification of the information objects on the metadata level is based on both name and subject authorities.

3.1 The role of National Authority files in making available digital content in Czechia

National Authority Files contain Personal name file (639,472 items), Corporate and Meeting name file (140,002), Uniform title (21,301), Topical term file (36,954), Geographical name file (27,170) and Genre/form name file (1,783).

National Authority files were originally developed with the aim to support the share cataloguing and the indexing of library materials. Name authorities were created according to AACR2[7] rules while subject authorities were originally created according to LCSH rules. Later, with advent of Word Wide Web, some new principles were introduced: post-coordination, the authorisation of isolated lexical units (not entire subject headings strings), the limited application of subject heading strings in bibliographic records. The main goal of authority records has been to establish authorised forms of names/appellations, their variants and evidence of these forms; they have been encoded using MARC formats.

3.1.1 Strengths of traditional authority files

In traditional authority files, standardized access points are established, they contain well documented best practices and are shared among libraries and other institutions as well. Libraries in Czechia participate in “Cooperative creation and use of national authority files”, while museums and galleries participate in “Museum Authorities” project and “Register of Fine Art collections” project. All these projects are based on national authorities; Museum authorities and Register of Fine Art collections are limited to the name (personal and corporate body) authorities.

3.1.2 Weaknesses of traditional authority files

Record based system

Traditional library authority files are record-based systems. They are created above all for humans, they are human understandable; but they are not available in machine understandable format, they are machine readable only.

Lack of data granularity

AACR2 and MARC format do not allow sufficient granularity of encoded information. MARC format doesn’t offer enough granularity in the data elements that should be clearly defined for Semantic Web and LOD purposes. “The MARC record was not created as a set of data elements but as a format for the storage and display of the text of library catalogue records” (Coyle, 2012).

[7] Since May, 2015 RDA Rules have been introduced.

Entities of the same type in different authority files

Another problem is that some entities are classified into various/different authority files in direct relation with MARC fields in which they are encoded. Some entities like events form part of two authority files: Meeting names file (conferences, seminars, workshops) and topical term file (wars, battles etc.). The named buildings form part of Corporate name file, the unnamed ones are classified in Topical file. These entities should be classified according to their attributes and ability to express the relationships to other entities rather than according to a MARC field.

Little information in authority records

Authority records contain only little information about the entity: they are concentrated on the preferred forms of entities name/appellation entered as main headings and their variants entered as see references. There was no possibility to enter all information needed for expressing properties/attributes and relationships of the entity in a structured way. On the contrary, it was possible to mention some information in note fields as a free text only.

Term describing type of corporate body

For corporate body headings, it is not possible to indicate a term describing type of corporate body (e.g.: ltd, plc), however for archivists, type of corporate body is one of the most important attributes that distinguishes an individual entity from other similar ones.

3.1.3 National Authority files applied in the INTERMI project

The national authorities were analysed in depth and it was decided to use them according to following principles:

- The standardized access points as preferred or variant forms of entities appellations shall be used.
- All information that provides context to entities in authority records shall be accepted in INTERMI system as well as entered in structured form.
- The tendency to make decisions based purely on the format (MARC format) should be avoided whenever possible.
- Proprietary conceptual model and data structure format have to be created to meet all specific needs of user´s communities.
- The INTERMI conceptual model shall be based on entities (not on authorities), it means there will be a strong tendency to move from the authorities to entities of value on the web.

3.2 Principles of object-oriented paradigm

Principles of object-oriented paradigm can be described as follows:

- **Data granularity** – the data is divided into smaller meaningful elements

which allow to define better type of data and to generalize them, so that the flexibility of data is supported.
- **Data inheritance** – the hierarchical structure of elements requires and supports inheritance of attributes or properties.
- **Data reuse** – the elements are designed to be reused in more than one location in the model and to enable further expansion of the model.
- **Polymorphism** – the elements can work as different elements depending on their use.

3.3 Relationships between entities in INTERMI conceptual model – event-centric approach

A relationship models an association between two or more entities to which all of the occurrences of those entities must conform. Frequently, a meaningful relationship exists between two different types of entity (e.g. creator – work, family – member of family, person – geographical object (place of birth). There are two types of relationships in INTERMI conceptual model: simple and complex.

To create a simple relationship in INTERMI knowledge model rules means to create a link to another entity (e.g. a person to person link) with possibility to specify role of linked entity (e.g. sibling, wife). It is possible to enter additional information about the relationship such as dates and note.

Some relationships are complex: expressing the complex relationships in its entirety requires to define and use one „descriptive“ event in the life or history of entity (e.g. birth, establishment, creation, destruction) and a set of relationships that are connected to.

Events represent specific actions that occur at a specific time and place and are encoded as information objects. We agree that “event modelling is so abstract that it can be used to describe cultural items and documentation of scientific observations” (Doerr, Kritsotaki, 2006).

There are two types of events in INTERMI conceptual model: events as class of entities and events that are used to express complex attributes and relationships by describing other entities, so called “descriptive entities”. Events (class of entities) include a named temporary event (long-term or short-term; one-time or repeated) and also named entities which are related to human activity such as culture, folk habits. “Descriptive entities” are connected to events in the life or history of entity (e.g. birth, establishment, creation, destruction).

Event-centric approach means that as many entity attributes as possible are described by event; we think that the event-centric approach reflects the concept of Linked Data.

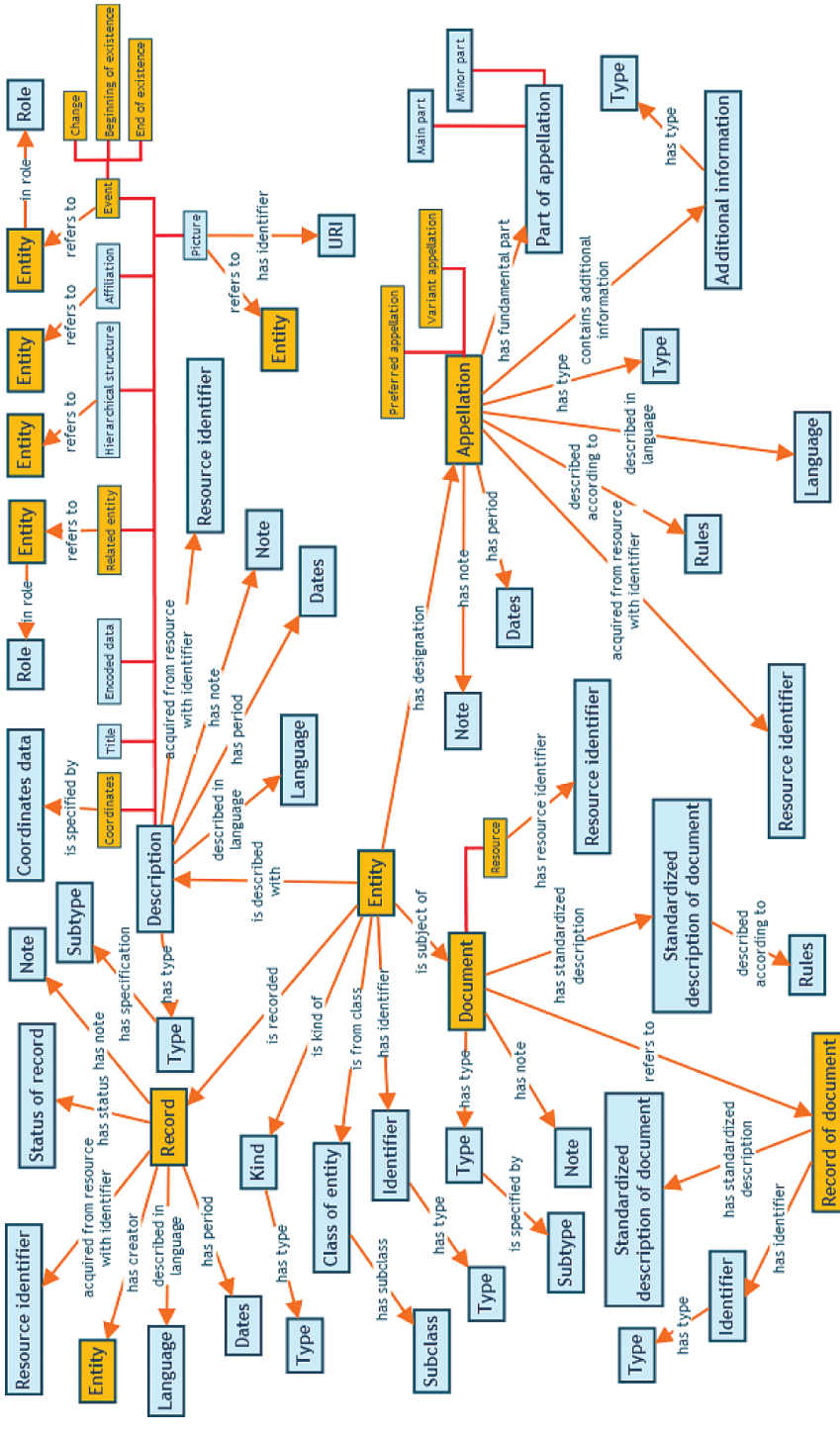


Figure 13: INTERMI conceptual model

3.4 Key elements of INTERMI conceptual model: entities and relationships

Entity is the most important element of INTERMI conceptual model.

Entities represent real-world identifiable objects. An entity is a concept (individual and general) that has an identifier, an entity type that describes the category of the entity and is defined by a set of characteristics which allow describing any entity used in the model. One of the most difficult tasks of the project was to define classes of entities and their characteristics. Based on an intensive interdisciplinary discussion it was decided that INTERMI conceptual model would be realized in 7 classes of entities: personal, family, corporate body, geographic object, work, event, and general concept entities. Each class is divided to subclasses that are meaningful subgroupings of its entities and need to be represented explicitly because of their significance to the INTERMI knowledge base application.

The definition of a class is based on common characteristics of the entities. The main goal of definition of classes was to easily identify and describe characteristics of entities belonging to specific class. However it is allowed to use characteristics from one class to describe entity primarily categorized in another class (examples are castles, dams etc.).

Table 2: Classes of entities and examples of subclasses of entities

Class of entities	Number of subclasses	Examples of subclasses of entities
Person/Creature	4	real persons, persons named in religious works, fictional and legendary persons, real or fictional non-human entities
Family	3	families, branches of the family, fictional families
Corporate body	12	territory with its own administration, political parties and movements, associations, firms …
Geographical object	9	states, countries, historical regions, geomorphological features
Events	3	organized actions and events (conferences, workshops), important days, wars, battles
Work	7	literary works, artworks, generally known documents (laws, constitutions), structures, named buildings (castles, churches…), programs, projects, grants
General concept	7	objects and their physical parts, categories and groups of not named persons and corporate bodies, abstract entities, materials, techniques

3.4.1 Brief characteristics of class of entities used in INTERMI project

Class of entities: Person/Creature

This class of entities includes real persons and persons named in religious works, fictional and legendary persons, and real or fictional non-human entities.

Person with more biographical and other identities

There are two possibilities to describe a complex person with more identities in INTERMI project. The first one (and preferred by INTERMI rules) is to describe the person with more identities in one record with one preferred appellation (more used, more known etc.), while other appellations are recorded as variant appellations. The second one mentioned in INTERMI rules is to describe the person with real and fictional identity separately – in two records. It seems to be useful when different biographical data used by alternative identity (as fictional curriculum vitae) exist. It is necessary to create relationships between all records connected to one entity.

Collective pseudonym, anonymous and unknown creator

A group using collective pseudonym is included in this class of entities if the name they use is similar to that of person. Specific entities as anonymous creator and unknown creator are included in the "person/creature" class of entities as well.

Class of entities: Family

"Family" class of entities includes families and parts (branches) of families. In the INTERMI project, it was decided to create separate class of these entities to describe specific characteristics of families in their complexity. As to this class, it is to be mentioned that different approach is applied in archives that concede a concept of families as corporate bodies (as named and organized group of people).

Class of entities: Corporate body

"Corporate body" class of entities was one of the most discussed classes regarding to different methods of description applied in archives and libraries. INTERMI rules adopt principles from archives and include the type of corporate body as one of attributes which determine the corporate body entity. Yet, this principle does not comply with the RDA rules. Separated record can be created for a part of corporate body.

It is important to differ when it is possible to classify the entity as corporate body, when as work (e.g. quarries, castles) or when as geographical object (e.g. village). Corporate body usually has seat, representatives, operating building etc.

Class of entities: Geographic object

Geographic object as entity is identified with stable set of geographic coordinates. The stages of development of geographic object are usually described in one record with variant appellations specified by dates indicating the time when the appellation was used. If necessary, INTERMI allows the application of separated record principle for every significant stage of entity development (with corresponding relations among records).

Class of entities: Work

This class represents the most complex class of entities in INTERMI project. The concept "work" within memory institutions communities varies greatly. Functional Requirements for Bibliographic Records (hereinafter FRBR) defines "work" as "a distinct intellectual or artistic creation. A work is an abstract entity; there is no single material object one can point to as the work. We recognize the work through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work" (IFLA, 2009, p. 17).

According to Cataloguing Cultural Objects "a work is a distinct intellectual or artistic creation limited primarily to objects and structures made by humans, including built works, visual art works, and cultural artefacts. …Works include architecture, landscape architecture, other built works, objects such as paintings, sculptures, murals, drawings, prints, photographs, furniture, ceramics, tools, costume, textiles, other decorative or utilitarian objects, or any other of thousands of types of artistic creations and other cultural remains. Performance art, installations, and site-specific works are included. Excluded are literary works, music, performing arts, language arts, culinary arts, science, religion, philosophy, and other intangible culture" (Cataloging Cultural Objects, 2006, p. 5).

As these are two very different approaches, it was very difficult to define this class of entities. In museums and galleries, the FRBR concept of "work" met with misunderstanding because of their different meaning of the concept (as "art work" – product of art creative power).

When identifying the class of entities "work" we define basic characteristics of the entity "work" as follows: "Work" is the result of intentional human activity, and it doesn't represent a process. Using this high level abstraction of description we are allowed to identify the class of entities "work" acceptable for all memory institutions communities. The description of entity "work" is concentrated to its current state. One work in its complexity is described in one record. If necessary, it is possible to create separated records on significant historical stage or part of work (with relations among records). "Work" class of entities consists of entities that have some material representation, but also of entities which do not exist in representation any more. This class include art and literary works, legal documents (laws etc.), standard documents, products, trademarks, games, grants, projects, buildings etc.

Class of entities: Event

"Event" class of entities includes a named temporary event (long-term or short-term; one-time or repeated) and also named entities which are related to human activity such as culture, folk habits.

According to INTERMI rules, every repetition or interpretation of event is described as separated entity. However, it is possible to create a collective (cover) record that consists of information related to complex history of events. This record is related to all repetitions/interpretations.

Class of entities: General concept

This class includes general concepts for specific entities such as categories of people, things, animals, plants, and also general concepts for abstract entities such as characteristics, scientific areas, art styles, etc.

3.5 Application of INTERMI conceptual model

General conceptual model was worked up into the INTERMI knowledge model rules, where basic rules and recommendations for specific properties registration are mentioned. Properties are connected to classes of entities. INTERMI knowledge model rules represent specific framework for general conceptual model and they were afterwards implemented to web interface for creating or modifying of entities.

General conceptual model is also transformed to proprietary XML schema used for data representation for web services. Data are sufficiently structured and it is possible to represent it in other various formats required by specific communities. It is necessary to keep in mind partial data reduction due to different scope of data representation in specific format.

3.5.1 INTERMI knowledge model rules

INTERMI entity record – structure of information about an entity

INTERMI entity is described/documented in record that contains a set of related data elements that are stored and processed together.

Table 3: Information includes in typical INTERMI entity record

Identifier	INTERMI ID – unique, obligatory; more identifiers are allowed, e.g. ID of National Authorities, Museums authorities, VIAF ID
Class/subclass	obligatory
Language	language in which the notes and additional information are entered – Czech language
Rules	used in descriptions of attributes and information about the entity, e.g. RDA, AACR2, CCO, Basic Rules (Archives), INTERMI rules

Dates	simple in the form of the text and complex (full information) which enables both: display and search
Notes	public and nonpublic
Brief characteristics, description, history	brief characteristics: maximum two sentences; description, history: maximum 12 000 signs
Appellation	preferred and variant forms, each form is identified by rules; maximum one preferred form for one rule
Parts of appellation	main part, another part of the appellation, additional part of appellation (topical, geographical, chronological information)
Classification	characteristics of the entity expressed through a topical term, category
Relationships and events used in description of the entity	simple relationship is expressed through link to other entity (with expression of the role of related entity, dates of relationship and note) complex relationship is expressed through events linked to more types of entities in different roles, dates and notes Three types of „descriptive“ entities are identified: beginning, end of the existence of the entity, change of the entity
Visual representation of the entity	e.g. picture
Resource for information about the entity	source of information is given to the whole record or to a specific property; standardized description is applied

INTERMI knowledge model rules provide instructions and guidelines on formulating data for memory institutions in Czechia. The aim was not to create proprietary INTERMI rule. At the same time, we wanted to keep the best practices and uniqueness of approach of memory institutions when it is required. We have carefully analysed the rules used in the memory institutions (AACR2, ISAAR (CPF), CCO, RDA) and we came to the conclusion that it was not possible to apply any existing rules in its entirety. We decided to apply RDA whenever possible. However, for specific areas of entities description, e.g. description of entity appellation, description of dates or definition of rules for mapping of controlled vocabularies used in memory institutions, specific instructions have been formulated. Specific instructions are complying with e.g. the Cataloguing Cultural Objects rules (e.g. description of dates) or Basic rules for Archives in Czechia (e.g. description of preferred and variant forms of corporate bodies).

INTERMI rules for entities appellation

Different memory community user groups in Czechia want to have different preferred forms for the appellations of the same entity (e.g. general object entity, corporate body entity); this idea is of key importance to the INTERMI rules: it allows creating more than one preferred appellation of entity in one entity record. However, it is necessary to assign each appellation to those

rules according to which of them it was created. This principle enables to provide multiple views on data of entity according to rules used in each memory institution and it is also helpful in thesaurus mapping process. We hope that the principle of multiple preferred appellation of an entity is main forthcoming solution, but in practical application the communities of memory institutions will be able to use one preferred appellation – because they will find it easy and reasonable.

INTERMI rules for dates in entity description

According to using dates in entity description we have decided to adopt principles of CCO. These principles are based on entering dates in two forms – for user access and for computer access. For user access, uncertain dates are described in text form (e.g. ca/approximately 1894), for computer access in format of ISO 8601, where the first date delimits potential initial date, the second one potential closing date (e.g. 1890–1900).

3.5.2 INTERMI rules for mapping of controlled vocabularies, registries etc. used in memory institutions

The aim of these INTERMI rules is to summarize methods used for mapping various vocabularies to INTERMI general object entities. We suppose that INTERMI controlled vocabularies mapping rules would be important for those institutions creating their own specific controlled vocabulary in a way that fits to Semantic Web.

Mapping/harmonization process depends on:

- the choice of a controlled system of terms that may serve as a basis for comparing and evaluating terms from other terminological resources,
- the choice of a tool for formal representation of terminological resources (SKOS),
- the description and characteristics (formal and semantic) of terms used in terminological resources,
- and the choice of methods applied in process of mapping/harmonization, such as lexical-based, concept-based and instance based-mapping.

The purpose of lexical-based mapping is to ascertain the degree of similarity of text strings.

Pojem	Identifikátor		
	Národní autority	EuroVoc	AGROVOC
politika	PSH8309	131558	6062
informatika	PSH6548	100223	3864
přírodní vědy	PSH11969	141614	4318
trh práce	PSH1264	142776	28713

Figure 14: Example of identical terms of three resources obtained by lexical-based method

Concept-based mapping is applied when terminology of specialized thesauri and controlled vocabularies is compared. It is realized as common graphs method or vector similarity method.

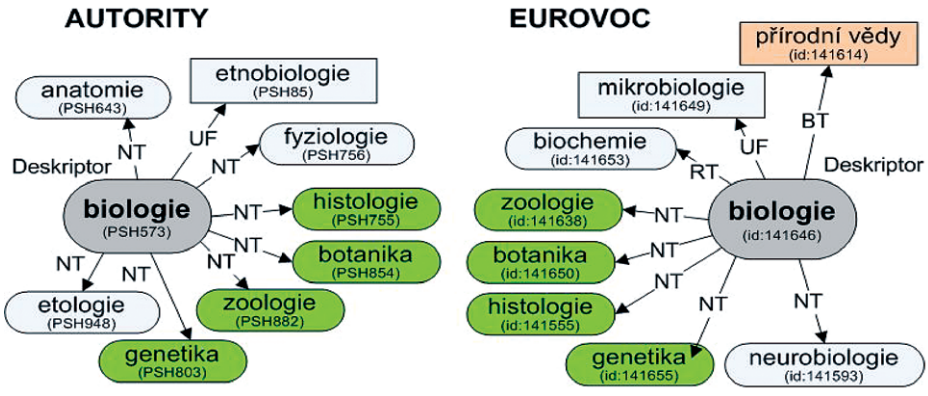


Figure 15: Example of Common graphs method – subgraphs of concept "biology" from terminological resources: National Authorities and EuroVoc

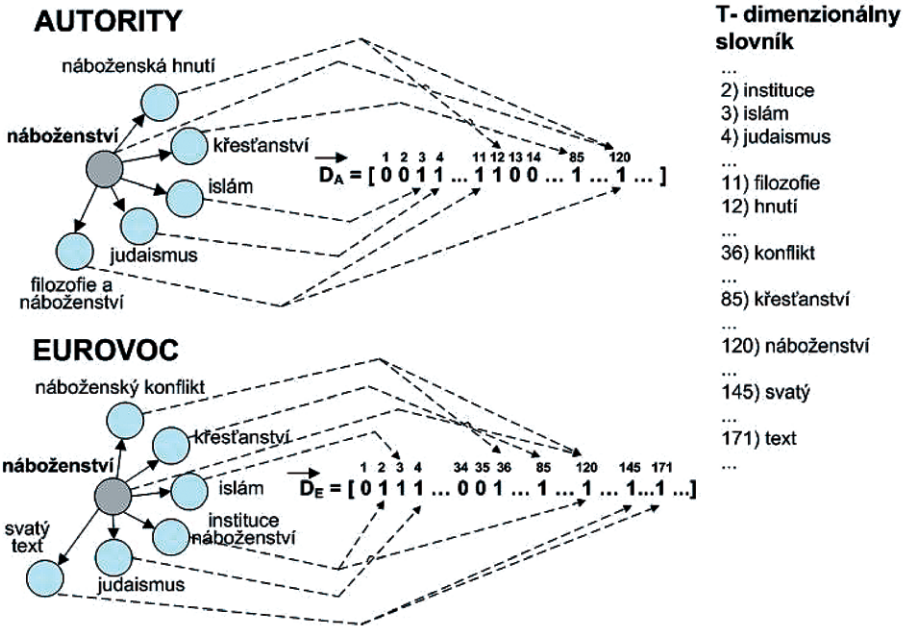


Figure 16: Example of Vector similarity method – T-dimensional vector of concept "religion" in terminological sources: National authorities and EuroVoc

When using the method of instance-based mapping, equivalent, broader, narrower, and related concepts based on the lexical and semantic similarities between source and target thesaurus are searched.

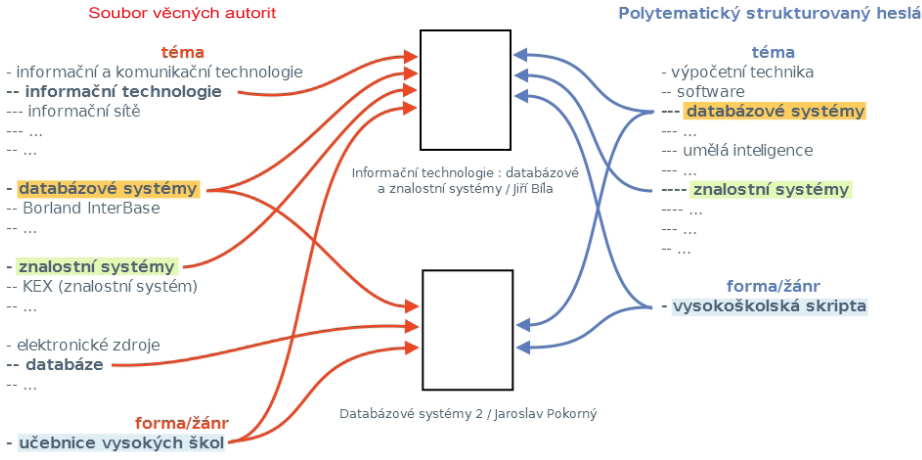


Figure 17: Example of instance-based mapping

The role of Topical authorities

Topical authority file has been chosen as a basis for mapping/harmonization of terminological resources used in memory institutions. Topical authority file represents a general set of terms with content covering all fields at general and specific level as well. Concerning Czech subject authority file, our intention is to create a tool enabling efficient mapping/harmonization, it means to transform it into a "general" ontology. It supposes:

- to define the meaning of topical terms exactly by adding qualifiers, UDC notations and scope notes, so that the topical terms are able to represent isolated well defined concepts,
- to express hierarchical structure exactly to be complete and consistent,
- to use permanent unique identifier which shall be language independent to avoid the necessity to change ID in conjunction with the change of preferred form applied in description of the concept,
- and to express topical terms file in machine-understandable way within the framework of the Semantic Web, when using RDF application SKOS; it enables data to be linked and merged with other RDF data by Semantic Web applications.

4 The model of cooperation and the plan of INTERMI knowledge database operation

The INTERMI project was not focused only on establishment of theoretical basis for knowledge database creation and development of software support, however it was focused also on creation of a cooperation model and a plan of knowledge database operation. A pilot plant was one of required results according to terms of project funding by NAKI programme. Pilot plant is defined by national information register of research and development results as: "verification of functionality [...] of procedures in larges scales, i.e. in a test or check operation, that serves to verification of properties, functions, susceptibility to failure and other monitored parameters important for application of new system into operation to maximal or planed output" (IS VaVaI, 2014, p. 4). Therefore the result has to be verified in experimental operation. In consequence of this condition it is necessary to create a cooperation model that would sustain filling and using of INTERMI knowledge database.

Model of cooperation includes specification of groups of users in INTERMI and also technological components.

4.1 Groups of users in INTERMI operation

Following levels of users are defined for INTERMI knowledge database operation:

- Processors – to insert and modify records by web interface for processing of entities; processors are incorporated to group subordinated to specific supervisor; groups can be created by domain, or affiliation.
- Supervisors – to supervise and approve drafts of new records of entities and modification of existing entities.
- Institutions – institutions contributing to knowledge database or using information from it by one of technology platforms (web services or Z39.50 server).
- Administrators – narrowly defined group of users that primarily manage account of users, assign access rights and incorporate users to groups subordinated to specific supervisor.
- Anonymous users – experts and public who use web interface for presentation of entities – i.e. searching and consulting INTERMI knowledge database.

Table 4: List of users' groups and they rights in INTERMI

	primary work space	insert of new records	modify of existing records	approve of modifications	manage accounts of users
administrators	web interface for administration	*no*	*no*	*no*	*yes* on all levels
supervisors	web interface for processing	*yes* within group	*yes* within group	*yes* within group	*no* (only own data for identification)
processors	web interface for processing	*yes* within group	*yes* within group	*no*	*no* (only own data for identification)
institutions	web services, Z39.50 server	*yes* within group	*yes* within group	*no*	*no* (only own data for identification)
anonymous users	web interface for presentation	*no*			

4.2 Technological components of INTERMI knowledge database operation

Technological components of INTERMI knowledge database are as follow:

- data storage and database system, application server and web server,
- web interface for data processing,
- web interface for data presentation,
- web interface for administration of users account and access rights,
- and technologies for communication with other systems – web services, Z39.50 server.

4.2.1 Data storage and database system, application server and web server

Data storage – several factors were considered during the selection:

- multidimensionality and more representation layers,
- performance,
- developers tools, platform, interoperability,
- connection to the next evolution,
- and hardware demands.

We considered not only the scope and resources of the project, but also the fact that we can re-use the existing solutions. It enabled to include number of extended functionalities into the base solution.

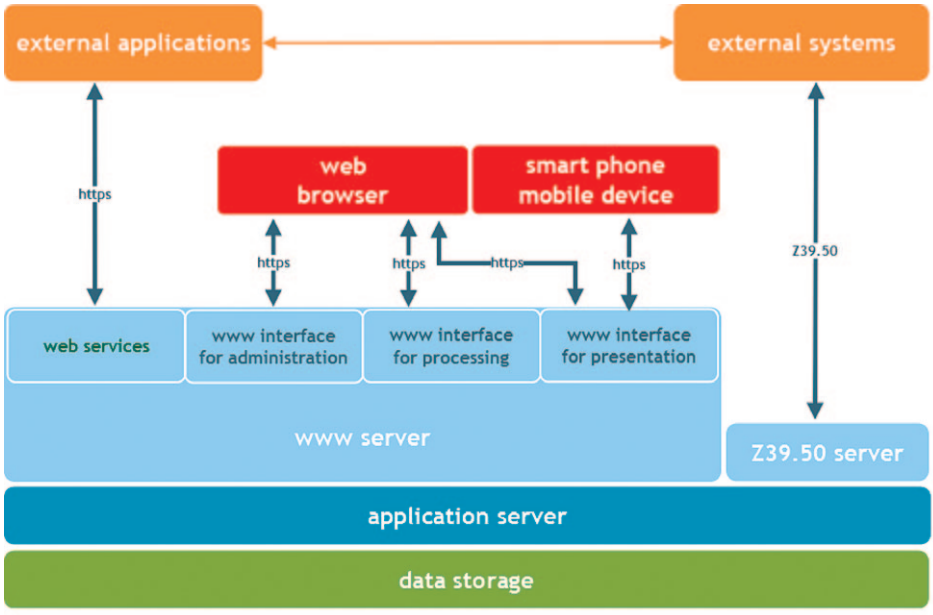


Figure 18: Architecture of system

We have opted for the data storage of InterSystems Caché®. This is a new generation of ultra-high-performance database technology. Caché is an advanced database management system with rapid development environment. It provides multiple modes of data access. Data is only described once in a single integrated data dictionary and is instantly available using object access, high-performance SQL, and powerful multidimensional access – all of which can simultaneously access the same data. This is a very big advantage and unique for Caché.

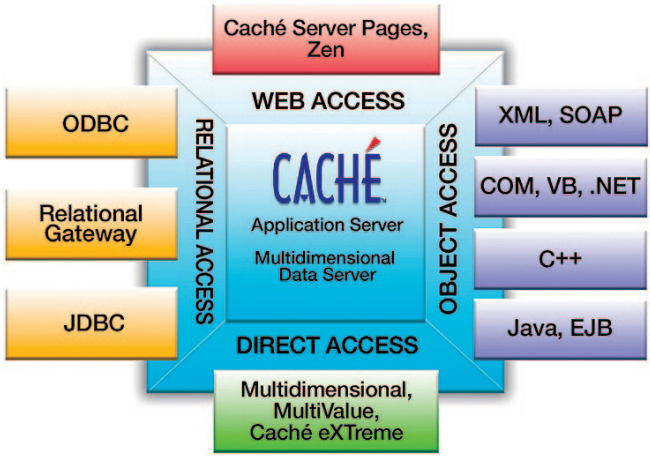


Figure 19: Caché's Multi-dimensional Data Server (InterSystems, 2015, p. 10)

Data model in Caché is based upon the ODMG (Object Database Management Group) standard. But the system also supports many advanced features, including multiple inheritance. Caché supports a full array of object programming concepts, including encapsulation, embedded objects, multiple inheritance, polymorphism, and collections.

The built-in Caché scripting languages directly manipulate these objects, and it also exposes Caché classes as Java, EJB, COM, .NET, and C++ classes. Caché classes can also be automatically enabled for XML and SOAP support by simply clicking a button in the Studio IDE. As a result, Caché objects are readily available to every commonly used object technology.

As it is standard in the object-oriented systems, the object is the basic structure of Caché and is an instance of class. In other ways class is the definitional structure and code provided by the programmer. It includes a description of the nature of the data (its "type") and how it is stored as well as all of the code, but it does not contain any data. An object is a particular "instance" of a class. For example, person who is defined by his ID="123456" and name "Tim Berners Lee" is an object of the "Person" class. (InterSystems, 2015, p. 7)

Object technology provides many benefits (InterSystems, 2015, p. 8–9):

- Objects support a richer data structure that more naturally describes real-world data.
- Programming is simpler – it is easier to keep track of what you are doing and what you are manipulating.
- Customized versions of classes can easily replace standard ones, making it easier to customize an application.
- The black box approach of encapsulation means programmers can improve the internal workings of objects without affecting the rest of the application.
- Objects provide a simple way to connect different technologies and different applications.
- Object technology is a natural match with graphical user interfaces.
- Many new tools assume object technology.
- Objects provide a good insulation between the user interface and the rest of the application. Thus, when it becomes necessary to adopt a new user interface technology (perhaps some currently unforeseen future technology), you can reuse most of your code.

We have extensively used one of the unique characteristics of Caché: the unlimited flexibility of the data storage. We namely used the database files ("globals") which are sparse multidimensional arrays. This type of storage together with variables and arrays are fully polymorphic, typeless entities that need not be declared or defined. They simply pop into existence as they are used and mold themselves to the data needs of what they are storing and how they are being used in an expression. Even arrays do not need any specification of size, dimension, type of subscripts, or data. For example, a developer might create an array called Person by simply setting:

```
set Person(„Novák“,“John“)=“I‘m a good person“
```

In this example, data was stored in a two-dimensional array using string data for subscripts. Other data nodes in this array might have a different number of dimensions and might intermix strings, integers, or other types of data for subscripts. For example, one might store data in:

```
abc(3)
abc(3,-45.6,"Yes")
abc("Count")
```

all in the same array.

Caché is fully object-enabled, providing all the power of object technology to developers of high-performance transaction processing applications. Besides providing inherently faster performance, it supports traditional indexes as well as bit-map and bit-slice indexes that can be used with real-time transactional data. For text searches, Caché enables word-aware searching in a number of languages.

We have also used the new indexing technology for our multidimensional data, which was introduced by InterSystems Caché – Transactional Bit-map Indexing – that leverages multidimensional data structures to eliminate these two problems. Updating these bit-maps is often faster than traditional indexes, and they utilize sophisticated compression techniques to radically reduce storage.

Caché also supports sophisticated "bit-slicing" techniques. The result is ultra fast bit-maps that can often be used to search millions of records in a fraction of a second on an online transaction-processing database. It also provides a powerful easy-to-use IDE for a rapid development of mobile and web applications, but with support for a variety of programming languages, it also gives you the option of using your favourite development tool.

But Caché is much more than a pure database technology. It includes an Application Server with advanced object programming capabilities, the ability to easily integrate with a wide variety of technologies, and an extremely high--performance runtime environment with unique data caching technology.

Caché uniquely offers lightning-fast performance, massive scalability, and robust reliability – with minimal maintenance and hardware requirements.

There are additional characteristics of the Caché product, which have considerable significance for our project. It serves as a base for integration platform Ensemble. It will enable lightning fast and easy integration and cooperation with outside systems, leveraging workflows, business processes modelling language and orchestration of third party web services.

There is also a DeepSee module of the Caché database product. DeepSee is embeddable software that makes it easy to enhance transactional applications to provide real-time analytics, giving users insights exactly when and where they need them. No data warehouse is required. It can be used in the scope of INTERMI to analyse relationships of the data, digging hidden patterns and thus reuse the existing data in the highly innovative way.

The Caché system thus not only addresses the current demands of the project, but it also opens it to future and paves the future ways.

4.2.2 Web interface for data processing

Technologies

The data process interface is provided as a web interface. It is accessible from any web browser with IE8 level or higher. From the technology point of view we opted for Ext JS by Sencha. The InterSystems ZEN technology was not mature enough in those times. Ext JS is the most comprehensive MVC/MVVM JavaScript framework for building feature-rich cross-platform web applications targeting desktops, tablets, and smartphones. Ext JS leverages HTML5 features on modern browsers while maintaining compatibility and functionality for legacy browsers. Ext JS features hundreds of high-performance UI widgets that are meticulously designed to fit the needs of the simplest as well as the most complex web applications. Ext JS templates and layout manager give us full control over our display irrespective of devices and screen sizes. An advanced charting package allows us to visualize large quantities of data. The framework includes a robust data package that can consume data from any backend data source. Ext JS also offers several out-of-the-box themes, and complete theming support that lets us build applications with our own layout and we could reflect brand of our project. Ext JS also includes an accessibility package (ARIA) to help with Section 508 compliance and with this future, our apps that could be usable for people who need assistive technologies such as screen readers to navigate the web. (Sencha, 2015)

Description of web interface for data processing

Created interface for data processing consists of main parts as follow:

- search of entities,
- form for editing,
- and list of open records.

Search of entities is designed as a search window that allows searching existing records using base search criteria (e.g. designation of entity, designation according to classes of entities, main part of designation etc.). Search criteria are improved by the lists of options for more precise search – mainly to retrieve unfinished records or records waiting for supervision. Search window is used at the beginning of processing; first of all to check if a specific entity is entered in the database, and then, when it is necessary, to search for related entity and to create a connection.

Forms for editing are designed for each class of entities; three forms are designed for class of entities work/creation, in order to specify distinctive properties connected to designation used in different groups of works/creations. Editing forms are divided to groups of properties. The groups of properties are created according to relation between properties or in the case when properties are following each other.

Table 5: Groups of properties in web interface for data processing

class of entities / forms	groups of properties in form for editing		
person / creature	forms of designation other identity short characteristics biographical data categorization	family / partner relationships activities awards other relations notes	pictures sources and authors other ID administrative data
corporation body	forms of designation history residence description and functions regulation	structure membership ownership categorization other relations	notes pictures sources and authors other ID administrative data
geographical object	forms of designation description history categorization coordinates / coded data	hierarchical structure membership other relations notes	pictures sources and authors other ID administrative data
family	forms of designation history and other short characteristics categorization	family relationships other relations notes pictures	sources and authors other ID administrative data
event	forms of designation place and organization history short characteristics	categorization other relations notes pictures	sources and authors other ID administrative data
artworks buildings other works	forms of designation description history position, location	ownership categorization other relations notes	pictures sources and authors other ID administrative data
general concept	forms of designation concept definition usage history	categorization coded data other relations notes	pictures sources and authors other ID administrative data

The chosen records are open on tabs of web interface, therefore it is possible to modify more than one record at the same time. List of open records is used for better orientation and allows closing records en masse.

Implementation of interface for processing as web interface has benefits mainly in its availability without installation of any specific software on user's working station. Requirement of stable Internet connection can be considered as weakness; it should be necessary in the case of desktop application as well because the solution is based on central data storage.

Some advices are summarized for minimizing problems when the Internet connection fails:

1 save data during the processing,
2 keep in mind that solution is based on session – session can be automatically closed in the case of inactivity or lose of connection,
3 it is possible to open more than one record, but think about it – more open records can cause mistakes in editing and in addition, open records are locked for editing by other users; close completed records – the use of interface will be more comfortable,
4 keep in mind security requirements – logout before closing the interface, as the session stay remains logged in for certain time.

Regarding to data processing, it is necessary to mention life cycle of entity record. Every new record or change is reviewed by supervisor. This process provides opportunity to detect especially formal mistakes (e.g. wrong application of INTERMI rules for appellation).

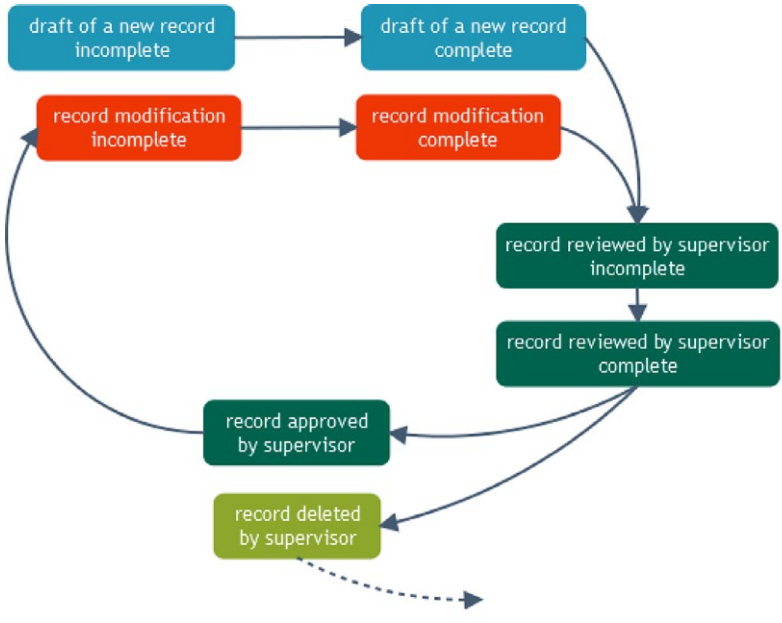


Figure 20: Entity record life cycle

4.2.3 Web interface for searching and data presentation

Technologies

The presentation and search layers were designed as an independent part. It uses the search engine built directly on top of the indexing machine and data storage. We strictly followed the principle of separation of application and presentation layer when designing the searching and presentation modules. Presentation layer is implemented in CSP – Caché server pages, which are a part of Caché. It enables tailoring of the look and the interfaces. We have used other technologies as well. The HTML mark-up language is a must for web pages development. We used also Microdata for marking the pages content enabling them for automatic processing. Systems like Google can add more user friendly data. As HTML cannot absorb all the complex information, we use also the WAI-ARIA (Web Accessibility Initiative – Accessible Rich Internet Applications). It enriches the semantics layer of the web pages and gives the information to assistant tools. It can:

- describe given widgets as “menu”, “tree item” or “slider”,
- describe the status of the widgets,
- describe the web page structure and its areas (landmarks) such as navigation, main body, header, footer etc.,
- handle the keyboard access for the respective web page objects.

The CSS – Cascading Style Sheets played substantial role, as well. It allows controlling the appearance of the page elements, as defined by HTML or XML mark-up languages. It leads us to another technology: LESS – Style sheet language, alias CSS pre-processor, which can generate the CSS for the browser, according to its syntax. It introduces programming language concepts into the CSS, such as variables, expressions or macros. It adds considerable effectiveness to the CSS. We used Framework Bootstrap for pages visual design and functionality of the search pages. It also ensures their transferability across the number of browsers, including the adaptation of the pages to the current wide variety of mobile devices and huge computer screens as well. We use the frontend framework Bootstrap on the pre-processor LESS layer. The jQuery library also transparently solves the browser’s incompatibilities.

On the client side we use JavaScript. As CSS isolate the presentation characteristics from the content in HTML, the jQuery JavaScript framework isolates the behaviour form the HTML structure. Last but not least we mention RequireJS, another JavaScript library used for optimized modular loading of the JavaScripts. We can thus link only one script into the page header. Only the RequireJS library and our configuration script are loaded together with DOM. All of the rest is loaded only once really needed.

Description of web interface for searching and data presentation

Web interface is designed for simple access to information about entities. There are two main ways how to retrieve information. The first one is to use

search box where it is possible to type query and choose search criteria (e.g. appellation, dates, places related to entities...). The second one is to browse classes of entities and related subclasses.

Displaying active links between entities is the most specific attribute of web presentation of information about entities. Links provide a way how to follow relationships among entities.

4.2.4 Web interface for administration of users account and access rights

Web interface for administration is used by very narrow set of users. Namely system administrators can manage the user accounts and their access rights to INTERMI database. From the technology point of view, the interface is identical as the web interface for data processing. Currently the rights are managed inside the INTERMI system; we have not addressed any rights sharing mechanism. This issue should be solved resort-wide but the unification process has not even started yet.

4.2.5 Technologies for communication with other systems – web services, Z39.50 server

Regarding to interoperability, the crucial role in INTERMI system lies in technologies for communication with external system. The first is protocol Z39.50 (also standard ISO 23950), that is well-known in libraries environment. It is Information Retrieval or Application Service Definition and Protocol Specification, ANSI/NISO Z39.50" – a protocol which specifies data structures and interchange rules that allow a client machine (called an "origin" in the standard) to search databases on a server machine (called a "target" in the standard) and retrieves records that are identified as a result of such a search (Lynch, 1997).

Libraries in Czechia create the database of National authorities using Z39.50 protocol since 2000, therefore it was important to keep this facility also in INTERMI project.

In the implementation of Z39.50 protocol, communication is initiated by Z39.50 client using IP address, port, database name (and user name and password if it is necessary). The set of attributes (named BIB1) is used for searching. For INTERMI project, we extend "use" attribute with new elements that contain specific information from entity record. Queries are built in PQF – Prefix Query Format, which is de facto standard for Z39.50. Within the INTERMI project we have implemented – in addition to the information retrieval functions – the entities creating and editing part.

Majority of other institutions (archives and museums), however, prefer web services for data interchange. Term web services often refers to an interface for a service oriented architecture (SOA), in which Web-based applications

dynamically interact with other Web applications using open standards that include XML running over HTTP, UDDI and SOAP. Such applications typically run behind the scenes, one program "talking to" another (server to server). (What are Web Services, 2015)

Web services communicate on the HTTP protocol basis. It is stateless on the contrary to Z39.50. Once server responded on the request, the connection to the client is closed. There are currently two main principles for the information exchange. Either SOAP – Simple Object Access Protocol or REST – Representational State Transfer. SOAP is an XML-based messaging protocol. It defines a set of rules for structuring messages that can be used for simple one-way messaging but is particularly useful for performing RPC-style (Remote Procedure Call) request-response dialogues.[8]

The third party systems can thus (via web services) not only search the data but create or edit the records in the INTERMI as well. The individual web services are described by WSDL – Web Service Description Language. It is XML based and it describes the calling syntax. So the calling system can read the information about the right way of the respective service usage and its formal content.

The need to cope with technological questions is self-evident, so we have to deal with two main problems:

- How to represent data covered by INTERMI conceptual model and to provide preservation of this data? This means to find internal data model for data storage that supports openness of INTERMI conceptual model, quick access to data and flexible presentation. We decided to use proprietary XML based structure that supports relations among entities.
- How to represent data for external systems (library and management systems, data management systems used in archives and museums, and the Web as well)? Concepts of Semantic Web and Linked Data represent suitable space for INTERMI because they provide technologies for presentation of data in context, not for presentation of content only.

Various standards related to Linked Data concept and Semantic Web were considered to find one or more suitable standards for data presentation. As to INTERMI project, following standards should be mentioned:

- SKOS – for representation of thesauri – can be applied on entities from class "General concept",
- FOAF – ontology for describing persons, activities and relations – can be applied on entities from class "Person/creature".

For maintaining continuity, it is necessary to provide data from INTERMI knowledge database in current structure that is frequently used in libraries – it means MARC21/XML for authorities.

[8] *(http://www.soapuser.com/basics1.html)*

Presentation of INTERMI data in several specific structures requires a selection of data that can be used in specific structure (or standard). This also requires more data transformation from internal to external structure.

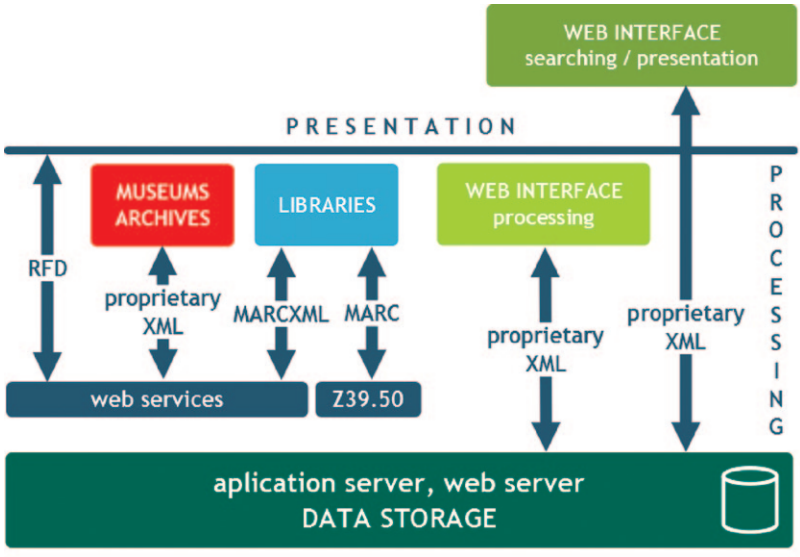


Figure 21: Technologies for data presentation used in INTERMI

XML – eXtensible Markup Language can be used as the document format. In this respect it belongs to the family of HTML, DOC, PDF. It can also easily transfer sets of complex objects, in much more flexible way then for example XSL, DBF or CSV. The XML marks have defined syntax, but there is no list of them. Thus in the data exchange there must be also the agreement on the meaning of the usage of the marks.

The XML was also used in conjunction with RDF – Resource Description Framework. It is a rather new way of information sources description. RDF can handle anything representable by subject predicate structure (transitive or intransitive). Thus we have the sources of information as subjects and statements about them. The simple statements, consisting of 1) subject, 2) predicate, and 3) object, are called triplets. The resource identification (the subject) is represented by URI – Uniform Resource Identifier.

The greatest practical problem in the resources identification is the universal uniqueness of the identifiers. In real world we fight with homonyms (one identifier points to two or more distinct things) or synonyms (one concrete thing has two or more identifiers – all used in given context). The INTERMI project intends bringing the solution for the existing problems in the area of cultural heritage. We are thus prepared to publish data in the RDF XML form, as well.

The last model, that we would like to touch here, is related to thesauri and/or other classification systems. It is SKOS – Simple Knowledge Organization System.

SKOS provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. As an application of the Resource Description Framework (RDF), SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes.

In basic SKOS, conceptual resources (concepts) are identified with URIs, labelled with strings in one or more natural languages, documented with various types of note, semantically related to each other in informal hierarchies and association networks, and aggregated into concept schemes.

In advanced SKOS, conceptual resources can be mapped across concept schemes and grouped into labelled or ordered collections. Relationships between concept labels can be specified. Finally, the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice or combined with other modelling vocabularies. (W3C, 2012)

5 Conclusion

Importance of the INTERMI project

INTERMI project provides a space for creation and preservation of data about entities used in memory institutions in Czechia. It represents an user-friendly tool for the access and share of the cultural heritage content across the memory institutions ensuring semantic interoperability at the conceptual level. The aim of the project is to create conceptual model and ontology for making Czech national cultural heritage content available on the web. The project builds on earlier research goals and projects, e.g. National authority files and Museum authorities, but in addition it aims for a comprehensive construction of an infrastructure for building a knowledge based model of the cultural heritage content and for the opening of its use in the form of working pilot operation. The project is complementary to the projects such as the National digital library, the National digital archive.

Impact of collaboration between memory institutions

It should be noted that the project has an impact on discussion among different communities from memory institutions. Experts have had opportunity to meet together and express their needs dealing with the identification and description of entities from their collections. At the end of INTERMI project we foresee an effective collaboration between professionals which would have a practical impact on their everyday practice when describing information resources using INTERMI entities. For realization of this goal, a set of INTERMI web services for application in local information system is provided.

Impact of user access to information

Using INTERMI entities in description of memory institutions collections will improve quality of user access to information on national heritage content published on World Wide Web; at the same time INTERMI project will demonstrate ways to present information about entities with applying technologies of semantic web.

References

American Library Association [et al.] (1978). Anglo-American Cataloguing Rules. 2nd Edition. Edited by Michael Gorman and Paul W. Winkler. Chicago : American Library Association, 1978.

ANDREJČÍKOVÁ, Nadežda [et al.] (2011). How to offer more to users. Pro-Inflow : Časopis pro informační vědy. 2011, (2): 82–97. Available from: *http://www.phil.muni.cz/journals/index.php/proinflow/article/view/833/965*.

AYLING, Peyton (2009). F25 Performance Plan F1 Work F15 Complex Work F14 Individual Work F2 Expression F22 Self Contained Expression R1 is logical successor of (has successor) [online]. 2009 [cit. 2015-10-17]. Available from: *http://slideplayer.com/slide/3334364/*.

BERNERS-LEE, Tim (2006). Linked Data [online]. 2006-07-27 [last modified 2009-06-18, cit. 2015-10-17]. Available from: *http://www.w3.org/DesignIssues/LinkedData.html*.

BOUNTOURI, L., C. PAPATHEODOROU, M. GERGATSOULIS (2010). Modelling the public sector information through CIDOC conceptual reference model. In: On the Move to Meaningful Internet Systems: OTM 2010 Workshops. Berlin : Springer, January 2010, pp. 404–413.

Cataloging Cultural Objects [online] (2006). Part one : General Guidelines. 2006 [cit. 2015-10-17]. 41 p. Available from: *http://cco.vrafoundation.org/downloads/PartOne_GeneralGuidelines.pdf*.

COYLE, Karen (2009). Metadata mix and match. Information Standards Quarterly. 2009, 21 (1): 9–11.

COYLE, Karen (2011). MARC21 as Data : a Start. Code{4}lib journal [online]. 2011-07-25, (14). Available from: *http://journal.code4lib.org/articles/5468*.

COYLE, Karen (2012). MARC elements [online]. [Last modified 2012, cit. 2015-10-17]. Available from: *http://futurelib.pbworks.com/w/page/29114548/MARC%20elements*.

COYLE, Karen (2015). FRBR, Twenty Years On. Cataloging and Classification Quarterly. 2015, 53 (3/4): 1–22.

COYLE, Karen, Diane HILLMANN (2007). Resource Description and Access (RDA). D-Lib Magazine [online]. January/February 2007, 13 (1/2). Available from: *http://www.dlib.org/dlib/january07/coyle/01coyle.html*.

Data Structure Standards (2013). Digitising contemporary art [online]. [last modified 2013-09-04, cit. 2015-10-17]. Available from: *http://wiki.dca-project.eu/index.php/Data_Structure_Standards*.

DOERR, M. [et al.] (2010). The europeana data model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. August 2010, pp. 10–15.

DOERR, M., A. KRITSOTAKI (2006). Documenting Events in Metadata [online]. 2006 [cit. 2015-10-17]. Available from: *http://www.cidoc-crm.org/docs/fin-paper.pdf*.

ELINGS, Mary W., Günter WAIBEL (2007). Metadata for All : Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums. First. 5 March 2007, 12 (3). Available from: *http://firstmonday.org/article/view/1628/1543*.

GAITANOU, P., M. GERGATSOULIS (2011). A semantic mapping of VRA Core 4.0 to the CIDOC Conceptual Reference Model. In: Metadata and Semantic Research. Berlin : Springer, 2011, pp. 387–399.

GERGATSOULIS, M. [et al.] (2010). Mapping cultural metadata schemas to CIDOC conceptual reference model. In: Artificial Intelligence : Theories, Models and Applications. Berlin : Springer, 2010, pp. 321–326.

Getty (2015). Categories for the Description of Works of Art [online]. [Last modified 2015-10-06, cit. 2015-10-17]. Available from: *http://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html*.

GILL, Tony (2004). Building semantic bridges between museums, libraries and archives : the CIDOC Conceptual Reference Model. First. 3 May 2004, 9 (5). Available from: *http://journals.uic.edu/ojs/index.php/fm/article/view/1145/1065*.

GUEGUEN, Gretchen [et al.] (2013). Toward an International Conceptual Model for Archival Description: A Preliminary Report from the International Council on Archives' Experts Group on Archival Description. The American Archivist. Fall/Winter 2013, 76(2): 566–582. Available from: *http://www.ica.org/download.php?id=3384*.

GUENTHER, R. (1997). The Application/MARC Content-type [online]. October 1997 [cit. 2015-10-14]. Available from: *http://whois.smartweb.cz/rfc/rfc2220/*.

HARPER, C. A. (2006). Encoding library of congress subject headings in SKOS: authority control for the semantic web. In: Proceedings of the 2006 DCMI International Conference on Dublin Core and Metadata Applications [online]. Available from: *http://dcpapers.dublincore.org/pubs/article/view/842*.

PARK, Hyoungjoo, Margaret E.I. KIPP (2014). Evaluation of Mappings from MARC to Linked Data. Advances in Classification Research Online. 2014 [cit. 2015-10-17]. Available from: *http://journals.lib.washington.edu/index.php/acro/article/view/14908/12495*.

ICOM (2014). The CIDOC Conceptual Reference Model [online]. [Last modified 2014-12-05, cit. 2015-10-17]. Available from: *http://www.cidoc-crm.org/*.

IFLA (2009). Functional requirements for bibliographic records : final report [online]. February 2009 [cit. 2015-10-17]. 137 p. Available from: *http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf*.

IFLA (2010). Functional Requirements for Subject Authority Data (FRSAD) : a Conceptual Model [online]. June 2010 [cit. 2015-10-17]. 75 p. Available from:

http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf.

InterSystems (2015). Caché Technology Guide [online]. [cit. 2015-10-17]. 52 p. Available from: *http://www.intersystems.com/assets/CacheTechnologyGuide.pdf*.

Introduction to the Semantic Web (2015). Cambridge Semantic. [cit. 2015-10-17]. Available from: *http://www.cambridgesemantics.com/semantic-university/introduction-semantic-web*.

IS VaVaI (2014). Definice druhů výsledků [online]. [cit. 2015-10-17]. 9 p. Available from: *http://www.vyzkum.cz/FrontClanek.aspx?idsekce=29415*.

JIN, Qiang (2012). Demystifying FRAD : Functional Requirements for Authority Data. Santa Barbara : Libraries Unlimited, 2012.

Joint Steering Commitee for Development of RDA (2009). RDA: Resource Description and Access : prospectus [online]. 1 July 2009 [cit. 2015-10-17]. 19 p. Available from: *http://www.rda-jsc.org/archivedsite/docs/5rda-prospectusrev7.pdf*.

Joint Steering Committee for Development of RDA (2010). RDA: Resource Description and Access. Chicago : American Library Association, 2010.

KAKALI, Constantia [et al.] (2007). Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies. In: Proceedings of the International Conference on Dublin Core and metadata Applications (DC-2007). August 2007, pp. 128–139.

KRAMER, R., R. NIKOLAI, C. HABECK (1997). Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. International Journal on Digital Libraries. 1997, (1): 122–131.

LANDRY, P. (2000). The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD). Presented at Classification and Indexing Workshop, 66th IFLA Council and General Conference, Meeting No. 181. Available from: *http://www.ifla.org/IV/ifla66/papers/165-181e.pdf*.

LE BOEUF, Patrick (2012). A Strange Model Named FRBRoo. Cataloging and Classification Quarterly. 2012, 50 (5/7).

LE BOEUF, Patrick (2015). A Basic Introduction to FRBRoo and PRESSoo [online]. Paper presented at: IFLA WLIC 2015 – Cape Town, South Africa in Session 207 – Cataloguing. [cit. 2015-10-17]. Available from: *http://library.ifla.org/id/eprint/1150*.

Library of Congress (2012). Bibliographic Framework as a Web of Data : Linked Data Model and Supporting Services [online]. Washington : Library of Congress, 2012 [cit. 2015-10-17]. 42 p. Available from: *http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf*.

LOURDI, Irene, Christos PAPATHEODOROU, Martin DOERR (2009). Semantic integration of collection description. D-Lib Magazine. 2009, 15 (7/8): 1082–9873.

LYNCH, Clifford A. (1997). The Z39.50 Information Retrieval Standard : Part 1: A Strategic View of Its Past, Present and Future. D-Lib Magazine [online]. April 1997 [cit. 2015-10-17]. Available from: *http://www.dlib.org/dlib/april97/04lynch.html*.

MITCHELL, E. T. (2013). Chapter 1: Metadata Developments in Libraries and Other Cultural Heritage Institutions. Library Technology Reports, 49 (5): 5–10.

RIVA, Pat, Martin DOERR, Maja ZUMER (2008). FRBRoo: enabling a common view of information from memory institutions [online]. Paper presented at: IFLA WLIC 2008 – Québec, Canada in Session 156 – Cataloguing. [cit. 2015-10-17]. Available from: *http://archive.ifla.org/IV/ifla74/papers/156-Riva_Doerr_Zumer-en.pdf*.

ROSE, Trish (2006). Cataloging in VRA Core 4.0 & CCO [online]. 2006 [cit. 2015-10-17]. Available from: *http://core.vraweb.org/pdfs/OLACpowerpt5_07.pdf*.

SCHERP, Ansgar [et al.] (2012). A core ontology on events for representing occurrences in the real world. Multimedia Tools and Applications. May 2012, 58 (2): 293–331.

SCHILLING, Virginia (2015). Introduction and Review of Linked Data for the Library Community, 2003–2011 [online]. [cit. 2015-10-17]. Available from: *http://www.ala.org/alcts/resources/org/cat/research/linked-data*.

Sencha (2015). Sencha Ext JS : overview [online]. [cit. 2015-10-17]. Available from: *https://www.sencha.com/products/extjs/#overview*.

SKINNER, J. (2014). Metadata in Archival and Cultural Heritage Settings: A Review of the Literature. Journal of Library Metadata. 2014, 14 (1): 52–68.

SOERGEL, D. (1995) The Art and Architecture Thesaurus (AAT): A critical appraisal. Visual Resources, X, pp. 369–400.

SQL (2015). Wikipedia: the free encyclopedia [online]. [last modified 2015-10-15, cit. 2015-10-17]. Available from: *https://en.wikipedia.org/wiki/SQL*.

SVENSSON, Lars G. (2013). Are Current Bibliographic Models Suitable for Integration with the Web. Information Standards Quarterly. 2013, (25): 7–12.

The MARC 21 Formats : Background and Principles [online] (1996). Revised November 1996 [cit. 2015-10-17]. Available from: *http://www.loc.gov/marc/96principl.html*.

TILLET, Barbara (2004). FRBR and Cataloging for the Future. Cataloging and Classification Quarterly. 2004, 39 (3/4).

Visual Resource Association (2015). 10 Key Concepts of CCO [online]. [cit. 2015-10-17]. Available from: *http://cco.vrafoundation.org/index.php/tenkey*.

W3C (2009). SKOS Simple Knowledge Organization System : reference. 2009-08-18 [cit. 2015-10-17]. Available from: *http://www.w3.org/TR/skos-reference/*.

W3C (2012). Semantic Web : Specification & Documentation [online]. [Last modified 2012-01-01, cit. 2015-10-17]. Available from: *http://www.w3.org/2004/02/skos/specs*.

WANNER, Michal (2005). Information system of Czech archives. INSAR. Summer 2005, (10): 2. Available from: *http://ec.europa.eu/archival-policy/docs/insar/insar10_en.pdf*.

What are Web Services? (2015). TutorialsPoint [online]. [cit. 2015-10-17]. Available from: *http://www.tutorialspoint.com/webservices/what_are_web_services.htm*.

ZUMER, Maja, Marcia Lei ZENG, Athena SALABA (2012). FRSAD : Conceptual Modeling of Aboutness. Santa Barbara : Libraries unlimited, 2012.

List of figures

List of tables

Facilitating Access to Cultural Heritage Content in Czechia
INTERMI project

Marie Balíková, Miroslav Kunt, Jana Šubová, Nadežda Andrejčíková

Graphic design: Richard Užák
Prepress preparation and typeset: Vojtěch Křesala
Proofreader: Bohuslava Králová

Published by the National Library of the Czech Republic
First edition
Prague 2015