

- [Mission](#)
- [Editorial Committee](#)
- [Process and Structure](#)
- [Code4Lib](#)

Issue 33, 2016-07-19

---

## **Checking the identity of entities by machine algorithms: the next step to the Hungarian National Namespace**

*The redundancy of entities coming from different sources caused problems during the building of the personal name authorities for the Petőfi Museum of Literature. It was a top priority to cleanse and unite classificatory records which have different data content but pertain to the same person without losing any data. As a first step in 2013, we found identities in approximately 80,000 name records so we merged the data content of these records. In the second phase a much more complicated algorithm had to be applied to show these identities. We cleansed the database by uniting approximately 36,000 records. The workflow for automatic detection of authority data tries to follow human intelligence. The database scripts normalize and examine about 20 kinds of data elements according to information about dates, localities, occupation and name variations. The result of creating pairs from the database authority records, as potential redundant elements, was a graph, which was condensed to a tree, by human efforts of the curators of the museum. With this, the limit of technological identification was reached. For further data cleansing human intelligence that can be assisted by computerized regular monitoring is needed, based upon the developed algorithm. As a result, the service containing about 620,000 authority name records will be an indispensable foundation to the establishment of the National Name Authorities. This article shows the work process of unification.*

By Zsolt Bánki (Petőfi Literary Museum), Tibor Mészáros (Petőfi Literary Museum), Márton Németh (Monguz Ltd.), András Simon (Monguz Ltd.)

## Background

---

The main topic of this article focuses on the semantic web related namespace building activities of the Petőfi Literary Museum. However, first we would like to give context to these efforts with a brief introduction to the first major semantic web related project in Hungary, at the National Széchényi Library.

The National Széchényi Library (NSZL) published its entire OPAC and Digital Library, and the corresponding authority data, as Linked Open Data in 2010 as one of the first public collections in Europe. The tools used were RDFDC (RDF Dublin Core) – MARC XML to RDF/XML conversion with XSLT for OPAC bibliographic data, FOAF for names, and SKOS for subject terms and geographic names. NSZL uses CoolURIs. Every resource has both RDF and HTML representations. The RDFDC, FOAF and SKOS statements are linked together. The name authority dataset is matched with the DBPedia (semantic version of Wikipedia) name files. NSZL also supports HTML link auto-discovery. All of the available linked data resources (names, subject authority, catalogue records) can be searched and retrieved to external resources in the semantic web via a SPARQL endpoint with specific browser tools (that are useful in machine-to machine communication). (Horváth, 2011b)

There was no specific project related to the semantic field within the library, although small developments pointed in the same direction. Three members of the directorate of informatics developed the semantic web based services when time permitted. In 2009 they realized that they had almost everything to be able to publish linked data on the semantic web: they converted the library thesaurus to SKOS format; via the LibriURI tools the OPAC records became accessible via URL (here is an example: <http://link.oszk.hu/libriurl.php?LN=en&DB=any&SRY=an&SRE=2616972>); and the URL-based search in the NSZL integrated system became available via the SRU protocol with the Yaz proxy tool. In addition, they could use the experiences of the LIBRIS semantic web implementation project. The NSZL main goals were: (1) Library datasets need to be open (to get data out), need to be linkable, and also need to provide links, and (2) Datasets must be part of the network, cannot be an end in themselves, and the system must allow hackability.

The major advantages of the RDF-based semantic model are the following:

- RDF Clients can look up every URI in the RDF graph over the Web to retrieve additional information
- Information from different sources merges naturally
- RDF links between data from different sources can be set
- Information expressed in different schema can be represented in a single model (Horváth, 2011a)

The project of NSZL, together with the efforts of the Petőfi Literary Museum, has become the major factor in planning the Hungarian National Namespace. In the following we will offer an overview of the related activities from the Petőfi Literary Museum. We will also offer a brief introduction into the concept of the Hungarian National Namespace.

## **A brief introduction to Petőfi Literary Museum**

Petőfi Literary Museum is the national museum for literary heritage. It is one of the most respected museums in Hungary. The history of the institute goes back to the beginning of the 20th century; it collects the tangible and intangible heritage of Hungarian writers. The museum is open to new trends and methods in the contemporary cultural heritage field and accordingly, IT technology has been present in a wide range of our activities since the 1980's. We have been working on the development of our collection databases for decades, and now service the entire research environment. Our integrated collection management system meets all contemporary standard international requirements and we use it for describing our collection items and to develop significant namespaces. The most important part of this system is the personal name database, which contains approximately 600,000 authority records: it is the largest such dataset in Hungary. This namespace contains controlled and certified data stemming mostly from printed resources, and it is published through the OPAC of the museum ([opac.pim.hu](http://opac.pim.hu)). The personal name space serves as the primary source for person name identification in the Hungarian library and museum system.

Our namespace includes different database genres (biography, bibliography, and genealogy). We developed several thematic databases for researchers which are unique in Hungary. Our colleagues working on the project seek to use only certified resources and have registered every difference among them, so the data are extremely useful for scholarly research and serve as a reliable reference.

Most of the databases based on lexicographical authority records are biographical, which presents information concerning the life and career of the person in question. For this project's method, a high percentage of the namespace does not have bibliographic records in the museum CMS system, yet they are valuable as a basis for processing collection items. This also means that the proportion of authority records connected to bibliographic items is constantly growing. (Kómár, Lengyel, & Simon, 2008)

## **The databases which we built in the namespace**

### **Hungarian Biographic Index**

This is the largest biographical database of our museum, one of the most important online sources for factographic information in Hungary used by librarians and researchers making bibliographies and bio-bibliographies. The records are not lexicon entries but consist of a series of data which makes identification of a person possible: date and place of birth, death, occupation, name variants. An important function of the Hungarian Biographic Index is that it compares and corrects the different data in its printed sources (lexica, documents, etc.) The Index defines certified data and, in case of birth and death datasets, it records the different data of the sources used.

## **Lexicon of Hungarian Immigrant Authors**

The term immigrant refers to authors and writers who left Hungary for a period of time, living and working abroad. The entries refer to this period of their career. Datasets include information on the higher education of the person in question, the date of departure from Hungary, the name of the destination country/countries, the states of the scholarly and/or public career, prizes and awards received, the institutes/societies in which s/he held membership, and the publication sources and the alias names used in them.

## **Awards and Awarded Personae**

The only Hungarian database of phaleristics consists of the names of awards and prizes of Hungarian relevance, made after 1945, as well as the names of the holders of these. The data collection is not confined to cultural awards but includes awards of other domains as well. The entries also contain the awarding entity, the place and date of the establishment of the award, and its history.

Our person name authority records do not solely consist of the requested normalized data elements, but further metadata such as occupation, and exact date and place of birth and death. Our rich datasets also include references to the data-source and, where applicable, the list of awarded prizes. We also indicate the relatives of the person in question. This data is available to be published on the semantic web.

To show the complexity of our namespace, we created a dataset of 30,000 authority personal name records which have 300,000 bibliographic record connections in the library catalogue of PIM.

## **Model of PIM authority records of person names**

Our records follow this architecture:

001 RecordID  
005 20160309083045.0  
090 \_\_ a Database  
100 1\_ a Personal name heading, Surname  
j Forename  
b Regal numeration  
g Title of nobility  
c Noble rank/Titles and other words associated with a name  
d Birth date/ Death date  
m Name suffix  
400 \_\_ a Name variant(s)  
d Name variant(s) date  
o Name variant explanation(pseudonym)  
t Note  
500 1\_ a Relationship: Personal name heading, Surname  
j Forename  
b Regnal number  
g Title of nobility  
c Noble rank/Titles and other words associated with a name  
d Birth date/ Death date  
m Name suffix  
o Relationship (mother, daughter)  
667 \_\_ a Occupation  
668 \_\_ b Occupation note  
678 \_\_ a Biography note  
856 \_\_ f Picture  
u URL  
900 \_\_ a Exact birth of date  
901 \_\_ a Date of baptism  
902 \_\_ a Place of birth  
9 Place of birth note  
904 \_\_ a County of birth  
905 \_\_ a Exact death of date  
906 \_\_ a Place of death  
9 Place of death note  
908 \_\_ a Country of death  
9 Country of death note  
909 \_\_ a Age of death  
910 \_\_ a Day of the funeral  
911 \_\_ a Cemetery  
912 \_\_ a Parcel  
921 \_\_ a Home  
922 \_\_ a Gender

923 \_\_ a Religion  
924 \_\_ a Walk of life date (marriage, degree, divorce)  
925 \_\_ a University education (years)  
926 \_\_ a Faculty  
927 \_\_ a Academic degree  
928 \_\_ a Membership  
930 \_\_ a Publication  
931 \_\_ a Translation  
940 \_\_ a Data source  
942 \_\_ a Birth data source  
943 \_\_ a Death data source  
960 \_\_ 4 Prize/Title type  
a Prize/Title  
d Prize/Title year

### **Collection Management System in the Petőfi Literary Museum**

The Qulto Integrated Collection Management System was originally written for library automation, especially for medium size special libraries. Research institutes, museums, and high schools have bought it for their libraries. Museums have begun to use it also as an automated inventory system. So its role has been to be not only a computerized catalogue system but to record and contain factographic information and additional data as well.

Libraries of Hungarian speaking people and in the areas of Central Europe where the Hungarian language is spoken and used (it is a bit bigger than the territory of the Hungarian republic) collect mostly Hungarian books, so Hungarian librarians are used to preparing form entries themselves, especially in special libraries. Therefore local cataloguing and indexing still has a significant role in Hungarian libraries nowadays. Hungarian library automation systems are completely compatible with systems of other countries, including the capacity for data exchange via Z39.50 or OAI clients. (Bánki, Lengyel, & Tóth, 2007)

The Qulto ICMS is based on a relational database that uses MARC as its internal format, and works with plenty of authority and quasi authority records, such as personal, corporate, and geographical names. However, copyright holders, suppliers, classification numbers, manufacturers, and prizes can also be authority data types. A couple of the fields are supported by dictionaries, so the main philosophy of Qulto is to help the librarian with additional information to prepare the most exact records and lessen mistakes as much as possible. The MARC structure with added quasi MARC fields is very segmented in the database and helps the user to tag the information. From these controlled data units all kinds of printed or electronic outputs or exchange formats can be easily prepared.

## **Data merge theory and practice**

---

### **Name authority: first step, basic concept of parsing**

Migrating the bibliographic and authority data from the former separately built databases to Qulto ICMS, we realized that several authority records were duplicates with different metadata content. The 620,000 name authority units held fewer records of living persons. So the “basic” personal record almost always duplicates data from the database of Graves of Hungarian Authors or from the database of the obituaries and death-notice as well. So in 2008, the first step of the authority name merge began.

#### **The primary record**

The basic name authority record is called “primary record”, which is one of the database units among the records describing the same person (human being). Doing the authority data merge, the name and date of birth (100\$a, 100\$j, 100\$d) were considered the basic data fields, and the other authority or bibliographic information in the 6xx or 9xx fields were considered the additional data. The records coming from different databases had different contents and all of them should have been inserted into the primary records, so all the information about the same persons and all the bibliographic records linked to them had to be merged into one fully enriched authority record, without any data loss. The merge of the first group of the records had been planned to be done automatically, without any manual intervention except for preparing the algorithm with which the identity of the authority records would be examined. Therefore very restricted conditions were set for the automatic identification. We wanted to allow only two records to be merged, if their identity was without any doubt.

#### **Priority of databases, defining primary records**

According to the richness of metadata, a priority order of databases was formed in order to choose the primary record from the redundant pairs. The primary record had already been enriched during the first phase of the authority process. Records had merged with metadata elements coming from the non-primary items. After needless records were deleted, the redundancy of the database decreased, and it became more exact and consistent, making it more suitable to fulfill the namespace function of including only one record for one person. By the first phase, 80,000 name authority entities were found redundant and had been deleted after the data merge. (Lill, 2012)

### **Name authority: second step**

After the first step it was clear that going forward we needed a new and more sophisticated parsing algorithm, which would have to be supported by human

knowledge. Following the data merging process new pairs had been created and in some groups of records anomalies were detected. Our aim was to automatically fix as many pairs as possible. By fixing the rest of the pairs as duplicates manually, lists and reports were created to help the librarians. Through this step those records were also reviewed that had not contained enough information for automatic identification.

## **Evaluation of the parsing**

For automatic fixing, an evaluation method had become necessary to rank the pairs. Concerning metadata richness and relevancy, pairs were evaluated with scores. Two authority records became a pair if their names were identical and their birth and death data weren't different. Using a strict score system, about 11,000 pairs returned as identical. A list was made from the pairs that had lower scores. The criteria for scoring the pairs were the following:

- Year of birth is filled and identical, 1 point
- Year of death is filled and identical, 1 point
- Both year of birth and death is filled and identical, additional 2 points
- Date of birth is identical, 2 points
- Location (village, town, city) of birth is identical, 2 points
- County of birth is identical, 1 point
- Date of death is identical, 2 points
- Location (village, town, city) of death is identical, 2 points
- Date of birth is not identical, -2 points
- Location (village, town, city) of birth is not identical, -2 points
- County of birth is not identical, -1 point
- Date of death is not identical, -2 points
- Location (village, town, city) of death is not identical, -2 points
- If one of the members of the pairs was the member of another pair where anomaly (minus point) was, it was marked as a suspicious item.
- All the 900 data pairs fixed as suspicious received, -2 more points
- If some birth or death data were missing, the pair received more negative points

- If Christian name was cut into more elements, the pair received more negative points

36,000 pairs were examined and 33,000 of them were fixed as identifying the same person. Common names (Tóth, Kovács, similar to Smith or Miller) existing in the database several times caused problems, especially if they were written in different ways (e.g. Kovács, Kováts, Kovách). There was a problem also with the Christian and family names having more than one unit, or having other name elements, titles, etc.

## **Operational problems**

---

There were pairs which contained more than one personal link to other records as husband, wife, father or mother, or had more than one birth or death date. These cases were checked manually.

As a part of the data merging process, additional data checking had become necessary and resulted in a more standardised form of the dataset.

- Geographical names existed in different forms or languages, or were written in different ways.
- The occupations also had to be unified, and subsequently a couple of pairs had to be fixed as identical.
- Some algorithms were defined in order to run automatic data merging processes to avoid future redundancy.
- Using experiences from the past, records could be listed where the person name was different but the other metadata elements were the same. If, from the four metadata indexes representing the year and place of death and birth, at least three are the same, the pair of authority records could be set as identical.
- Sometimes authority records having too few metadata elements could be identified by the bibliographic records linked to them.

The aim was to prepare full name authority records, including all the necessary information (school, prizes, family connections, death-reports, etc.), concerning a person in a database record.

The aim of the parsing algorithm was to fix duplicated authority records concerning the same person. However, these records could take parts of other pairs as well. If we set a record as primary, it could take part in another pair as a non-primary element as well. We had to take care to avoid creating networks of pairs that resulted

in a loop, where each of them appears once as primary and once as non-primary item. This had been avoided by the way we precisely ranked the databases, and set the records as primary using a strict and uniform algorithm.

In cases where the two items of a pair came from different databases the one from a higher-ranked database got the primary rank. In other cases the primary status was chosen according to the record ID. The richness of metadata wasn't checked in this step, because all metadata elements had been merged. If a primary record in a pair had parsed in another case as a non-primary one, its primary pair had been placed into the pair in which it existed as primary, so every pair was completed with a super primary unit. Super primary members of groups could be only those records which had never been labelled as a non-primary item.

Another critical element of the parsing method was the problem of "forks": networks of pairs which have more than one super primary element. These pairs were always checked manually. This problem could occur in cases of missing metadata elements. In most cases the date of death was missing.

Through checking the result of the data merge we received useful information about the database as a whole. In some cases, we could correct the database records which hadn't been part of the parsing process. After the data merge all of the work was checked by randomly selecting items.

### **Technical elements of the steps of parsing**

The first step was the selection of the authority records from the database to be part the merging process. The valid record had to be in non-deleted status, and had to contain some data elements, for example, Christian name and year of birth. The selected names were normalized, so that the variations of some letters, such as "y" and "i" or "ts" and "cs", were unified, and all the punctuation was eliminated. From Christian names having more than one unit, control strings were created containing all the combinations of the Family name and other name segments for the Christian name. For example: from this string "Mezőhegyesi Szilveszter Aladár", three results were generated:

Mezőhegyesi Szilveszter Aladár  
Mezőhegyesi Szilveszter  
Mezőhegyesi Aladár

From the index data of Family and Christian names, year of birth and death, a key string was generated on which parsing of authority records was based. In cases of control subfields (exact date of birth and death, location of birth and death, occupation), additional name elements were added to the key strings. By doing this,

the fully segmented Quito MARC based relational database records were built into strings for parsing and control.

Through a detailed analysis of the pairs, we found anomalies (differences according to names and locations). We marked the pairs with numbers, to help the librarian to determine the identity of the primary and non-primary record. The correct items of information increased, and all the anomalies decreased during this process. Through an analysis of this list we could decide which pair joined to two identical records, defining the same person in this way. We could determine also if the members of the pairs were identical without manual intervention.

Merging was done in several steps. At first those pairs were merged which were without any doubt the same. Secondly, from the dubious pairs a stack was chosen to merge without human intervention. By the third step, the rest of the list was analysed one by one by the staff of the museum.

### **Merging process step by step**

1. database backup
2. creating the key string of the record for the comparison
3. normalizing
4. creating the pairs from the data heap
5. creating graphs as a network of pairs (one record could be a member of several pairs)
6. erasing loops and forks
7. creating the pairs that really could be identical
8. erase the duplicates of pairs (these duplicates were created through the splitting of Christian names; some records could be involved in the parsing process several times)
9. adding super primary records to the pairs
10. control of the pairs, removing those with a deleted element (the whole operation had been made in a dynamically changing catalogue environment, and for this reason it needed fairly significant time)
11. revision of parsing by manually checking hundreds of pairs
12. preparing the list of checked pairs
13. Through the data merge as a process, the first step is the creation of a view dataset for data-checking on the ICMS interface. The next step is the real data

merge, by moving the data subfields from the non-primary to the primary record. If the non-primary record had a new data field, the new content was copied into the primary record. All the other data elements of the non-primary record remained on the view dataset.

14. deleting the duplicates of data subfields and authority records

### **Further plans with the PIM namespace- The conception of the Hungarian National Namespace**

The Hungarian National Digital Archive, National Széchényi Library, Hungarian National Archive, and Petőfi Literary Museum started a broad collaboration project in 2012 with the national namespace concept. The idea of the concept is to focus on the experiences of the process of building large collaborative namespaces as professional service frameworks created by international collaboration (GETTY, VIAF, ICONCLASS). It has become clear for the members of the project that the focus of future professional activities will be on the usage of namespaces in the semantic web environment. The target groups are both the institutions with their professional needs and the users of the public collections in the broadest sense. Professionals from different institutions are agreed that the institutional namespaces and other high-quality semantic-web compatible metadata collections will form the basis of the common Hungarian namespace. Following this conception, cleaning the dataset of names, including the identification and data clearance of entities, has been a major contribution by the Petőfi Literary Museum towards the creation of the Hungarian national namespace.

The creation process of the national namespace is still in progress. However, the concept is already available. The collections from the different institutions' collections can be fully connected and are interoperable with each other. Using joint namespaces seems to be the best way to represent the Hungarian cultural heritage as a whole, describing all of the available contexts and connections through the different segments of this broad area. All the cultural databases should be available through a common interface regardless of the traditional institutional barriers among the archive, library and museum sectors. Collections and datasets being stored by the individual institutions can be represented in their own local contexts. On the other hand, the public can have access to the digital cultural heritage as a whole. The content of the local collections could be represented through a framework of an integrated system (based on namespaces with common standards on the semantic web). The primary user group of the future national namespace will be the general public. Through the first step of the implementation process, however, the collaborating institutions must focus on their own institutional contexts and needs in order to create their own high quality datasets, service workflows, and infrastructure. The second step in the future should focus on the implementation of all the services

and functionalities focusing on the general public. Another strategic goal of the Hungarian National namespace can be a data authority function. The National Namespace must serve qualified, controlled, and clear data on request to anyone, whether a public collection and other public institutions, or individuals.

## **Conclusion**

---

Through this article our aim was to offer a brief overview of institutional efforts from the Hungarian cultural heritage sphere in the field of semantic namespace building. Our primary example was the namespace project of the Petőfi Literary Museum, with a brief description of efforts in the Hungarian National Library as well. The recently acquired experiences, skills and competencies through local efforts have contributed to making the Hungarian National Namespace. Our main hope is that the project can perhaps be realized to its full extent in the near future. The local institutional projects could be amalgamated into a successful national service model. Following its completion, we hope that it will also be recognized by the general public in Hungary by representing the Hungarian Cultural Heritage in a complete way that has never been available before.

## **References**

---

- Bánki, Z., Lengyel, M., & Tóth, K. (2007). "Ablak" a múzeumokra: A Petőfi Irodalmi Múzeum speciális adattárai a Huntékában. Networkshop 2007. Eger: NIIF. Retrieved 26 May, 2016, from <https://nws.niif.hu/ncd2007/docs/aen/095.pdf>
- Horváth, Á. (2011a). Linked Data at NSZL. Retrieved 26 May, 2016, from [http://nektar.oszk.hu/w/images/0/04/LinkedDataAtNszl\\_06.pdf](http://nektar.oszk.hu/w/images/0/04/LinkedDataAtNszl_06.pdf)
- Horváth, Á. (2011b). National Széchényi Library Semantic Web wiki. Retrieved 26 May, 2016, from [http://nektar.oszk.hu/wiki/Semantic\\_web](http://nektar.oszk.hu/wiki/Semantic_web)
- Kómár, É., Lengyel, M., & Simon, A. (2008). A Petőfi Irodalmi Múzeum személynév állományának egységesítése és szűrése: Hozzájárulás az NDA névtér projektjéhez. Dunaújváros: NIIF. Retrieved 26 May, 2016, from <https://nws.niif.hu/ncd2008/docs/aen/046.htm>
- Lill, J. M. (2012). Kontrolliertes Vokabular. Wieso? Weshalb? Warum? Vortragsfolien vom 13. MusIS-Nutzertreffen am 18. Juni 2012 in Mannheim. Retrieved 26 May 2016, from <https://swop.bsz-bw.de/frontdoor/index/index/docId/850>

## **About the Authors**

---

Zsolt Bánki received a college degree in Library Science and a degree qualifying as a school teacher in Hungarian Language and Literature from Dániel Berzsenyi Teacher Training College, Szombathely, Hungary. He graduated as a teacher of roman catholic religion at Péter Pázmány Catholic University, Budapest, and obtained a Master's level degree in Library and Information Science from Loránd Eötvös University, Budapest. He started his library career in the National Library of Foreign Literature, then worked for the National Széchényi Library, Budapest. His professional interest is focused on systems of museum informatics and professional digital museology standards, theoretical issues on digitization, and planning and implementing national and international databases. He has been the Director of Informatics in Petőfi Literary Museum, Budapest, since 2006. He has been involved in several national and international projects and is a leading member of several Hungarian professional organisations.

Tibor Mészáros received a college degree in Library Science and a degree qualifying as a school teacher in Hungarian Language and Literature from Dániel Berzsenyi Teacher Training College, Szombathely, Hungary. He graduated as a teacher of roman catholic religion at Péter Pázmány Catholic University, Budapest, Hungary, and has been a librarian at the Petőfi Literary Museum since 1988. He has been the editor of the Hungarian Biography Index since 2010. A primary focus during his career has been on the literary life of the famous Hungarian writer, Sándor Márai. He is an author of several books, editor of bibliographies, curator of exhibitions, and also a professional lecturer related the work of this renowned writer.

Márton Németh graduated from the master's-level university program in history at the Faculty of Arts, University of Szeged, with high-school teacher-training specialization. He subsequently completed his master's degree in library and information science at Szeged. He also holds a master of social science degree in European studies from Aalborg University, Denmark, and recently graduated from the International Master of Social Science in Digital Library Learning offered jointly by Oslo University College, Parma University, and Tallinn University. He has worked as an IT librarian in the Directorate of Informatics, National Széchényi Library, and is now a digital information specialist at Monguz Ltd. in Budapest. He is currently working on his PhD at the Doctoral School of Informatics, University of Debrecen, Hungary, focusing on the interrelation of public collections and the semantic web.

Simon András graduated from Eötvös Loránd Tudományegyetem Budapest Faculty of Arts History – Librarianship, and did postgraduate work at the Economic University of Budapest as Postgraduate Economist of faculty of Marketing. He worked as system librarian in the Economic University of Budapest, and at the Magyar Országos Közös Katalógus (Hungarian National shared Cataloguing System) as system manager. He has been on the MTA SZTAKI ITAK Library automation team, and Monguz Ltd., working on library and museum automation, taking part in

projects, planning, developing, preparing adaptations, educating, selling and supporting content management systems.