STRUCTURE AND SEMANTICS, COHERENCE AND NETWORKS – THE LIVING BIBLIOGRAPHIC UNIVERSE

REFLECTIONS OF A CATALOGUE LOVER IN HONOUR OF MIRNA WILLER, A DATA SCIENTIST[1]

Claudia Fabian

Department of Manuscripts and Rare Books, Bayerische Staatsbibliothek, München, Germany

Keywords:
European printed book (1450-1830), Heritage of the Printed Book (HPB), data management, data science

Abstract

Remembering our common involvement in the Consortium of European Research Libraries and, in particular, the building of the HPB (now the Heritage of the Printed Book database), these reflections on the "living bibliographic universe" follow the development of the catalogue and the structure of records from a monolithic approach towards a modular, semantically rich network. From the point of view of a non-specialist but at the same time a catalogue lover, the importance of data management and data science in contributing to the great vision of the bibliographic universe are illustrated in ways that are easily understood by all those who have to deal with issues of this kind. The HPB, which gives access to the early European printed book (1450-1830), is considered here as a "microcosmos", enabling general issues of the bibliographic universe to be grasped.

1 I thank all who helped and advised me in writing this paper: Gabriele Meßmer, Ann Matheson, Marian Lefferts and Maike Kittelmann. Responsibility for any errors is the author's.

Mirna Willer and I first met in 1994[2] in the Consortium of European Research Libraries (CERL), which celebrates its 25th anniversary in 2019. We shared a common enthusiasm for conceiving and building a European database, a "specialized bibliographic universe". The Hand Press Book database, which was renamed the Heritage of the Printed Book database (hereafter HPB) in 2008, is a central bibliographic database integrating, maintaining and updating records for early printed European books (1450 to 1830) from different sources in a common structure. Active since 1997, the database was initially only accessible to members of the Consortium and their associates: in January 2018, the much extended and steadily growing database became freely accessible to all. Today, the database comprises more than 7.8 million records.

The database is not only an extremely useful tool for identifying and locating early printed books, and sharing and reusing records: it is also a testbed putting into practice concepts about international bibliographic control, format convergence and record sharing. Records come from different countries and from different institutions within these countries. They have been created using different cataloguing codes, rules and data formats, and from both book-in-hand cataloguing and the conversion of printed or card catalogues. As a result, the database also contains historical rule sets translated into modern data structures. Since the start, full well-structured records have co-existed with very short, even deficient, and only roughly structured records, elaborate national retrospective bibliographic files with simple censuses of copies. Some records derive from local cataloguing and specialist projects while others are from vast union catalogue files with different contributors. Some have detailed copy specific information, especially provenance, on which CERL is particularly keen, as is illustrated by the setting up of a dedicated working group from 2003, the "Can You Help" initiative on provenance since 2006, and the opening of a CERL Provenance Digital Archive in 2018. A number of European languages and scripts can be found in HPB records: this is particularly challenging in note fields.[3] We can also see different forms of standardization and spelling traditions. Last but not least, the individual character of early printed books – in their construction and in their tradition and lifecycle up to today's updating of information and adding links to digital copies of the originals – contributes to multiplying the differences between these records.

2 And particularly CERL's Advisory Task Group (ATG), established in 1994, in which Mirna Willer played an important role, first as a consultant, member, then as Chair from 1999 to 2007. In 2012 the ATG ceased its activities.

3 E.g. "Painovuosi nimekkeestä. - Arkit: 3 arkintunnuksetonta lehteä, L-M4", or "Titelblad in meer kleuren". But specialists in book history understand.

Some records have fingerprints based on different systems,[4] transcription of title-pages with line breaks, standardized publication information, detailed content description, differentiated language or typography details and citations for reference works while others have short, differently or even unmarked abbreviated or standardized titles according to modern spelling. The treatment of multi-volume publications, periodicals, series and the analysis of content (an area to be further developed in the future) differs substantially between different files. Microform publications[5] and now digitized copies (sometimes described in a new, separate record, and sometimes with only a link to the digital copy added to the record for the original) have to be taken into account – thus opening the way from a bibliographic database to a digital library. The HPB is a living and growing bibliographic universe adapting to the endeavours of updating and modernizing information about early European printing.

All these records co-exist in the HPB and we have to consider this as an asset, not only for researchers in the rich field of historical imprints, the primary intended audience, but also for catalogue lovers and data scientists. There have been continuing discussions about the clustering of records for the same manifestation.[6] The well-known and evident difficulties of identifying duplicate records have so far stood in the way. This makes the HPB a wonderful comprehensive showcase not only for the collections it describes but also for the diversity and richness of different cataloguing traditions. The aim of the HPB is not to simplify or unify this richness but to conserve it as far as possible in its completeness in a controlled, authoritative, trustworthy and well-developed format-based environment. This allows a structured and relatively reliable access to the flourishing "jungle of diversity", which addresses the bibliographic needs of a research-focused user and a specialist cataloguer more than appealing to the general public. Care continues to be taken for the enrichment of this rich central database and for its presentation and maintenance.

In this the HPB positively differs from OCLC's WorldCat and from USTC (the Universal Short Title Catalogue, "a collective database of all books published in Europe between the invention of printing and the end of the sixteenth century").[7] Both merge records according to unrevealed parameters: USTC is a research-de-

4 Perfect, if clearly mentioned, e.g. "The fingerprint identifier follows the rules used by the STCN."

5 Microfilm records were never very much in the focus of CERL's own activity as they were well cared for by EROMM, accessible, e.g. via the CERL portal.

6 The term "manifestation" is used in RDA terminology.

7 Only in 2007 did its scope reach beyond French vernacular imprints, which were catalogued book-in-hand in several libraries. Some of the bibliographic files later integrated into USTC are identical to those integrated into HPB, among them ISTC, VD 16, STCN.

veloped tool integrating library resources while HPB is firmly anchored in the rich library service-oriented structures which provide a reliable infrastructure.

The HPB also illustrates the opportunities, challenges and limits of networking and transferring the cataloguing tradition from an old fixed and static constellation to individual files and then into a consolidated ever more open bibliographic universe. Like our human universe, the bibliographic universe is a living, dynamic, modular, multiple, moving, and interactive and interdependent network. In this way, the HPB is a landmark, a bibliographic "microcosmos" for a clearly defined entity: the European printed book from 1450 to 1830. Looking into the HPB, there are some uncertainties around this focus: modern secondary formats for historic prints, some printed music (which ought to be excluded, as there is a suitable material-specific central international environment within RISM[8]) and maps (where decisions about inclusion or exclusion continue, and where no central material-specific database is in place). There are also records that are difficult to identify and omit: the long tail of errors in format, cataloguing, data preparation and data analysis. Indeed, one of the most useful and important services CERL offers those who deliver their files and updates for inclusion in the HPB is a careful analysis of their file, allowing the identification of these and other errors. They can be detected by means of automated quality checking, which data transfer enables: undetected they may interfere with the usability and retrievability of data.

Let us now follow the history of cataloguing, which is so well translated and conserved in the HPB, and observe the ongoing process of transformation, modularisation[9] and semantic enrichment on the way towards the semantic universe.[10]

Monolithic catalogues of the past – "when Mirna started librarianship"

We remember areas in our libraries where the old card catalogue sat like a huge cupboard or a safe.[11] The catalogue as a whole and each record within it was locally fixed, sometimes clearly and certainly firmly and irreversibly structured. Corrections, additions and expansions were a problem and a threat. The

8 Répertoire international des sources musicales, the RISM-OPAC is an online catalogue of musical sources, both printed and handwritten.

9 I use this very general term to define the division of a system into its different elements, components, for structure as well as content.

10 For a fuller, better informed, specialized and reliable approach to all the issues raised in this paper, cf. Willer and Dunsire (2013). The author of this paper apologizes for any inaccuracies and inconsistencies in comparison with this monograph. Had the publication been seen earlier, this paper might not have been written. The term "semantic universe" used here means the semantic web.

11 A suggestion for an interesting read: Barbier, Dubois and Sordet (2015).

sequence of cards or entries permitted pre-coordinated searches. It was a "local star", the heart of the library, and in its content already internationally harmonized. ISBD, which IFLA has overseen for decades, allowed completeness of information and a reliable structure in the area-focused presentation and sequence of the bibliographic elements. This enabled the copying of catalogue cards as the first approach to reuse and exchange data. These structures – defining areas, relying on sequence and description marks – were more of a professional discipline, although they were meant to reliably inform and guide the international user. And there were already connections created to the outside world: ISBN and ISSN at an international level, LC or national bibliography numbers and, of course, the shelfmark, pointing towards the real book on the shelf.

International concern for the harmonization of cataloguing rules started after the Second World War,[12] driven by a spirit of efficiency. International care for the reuse of information was primarily oriented towards the modern book. Research libraries worldwide built collections quite similar to one another in their aim to best serve the needs of their users. However, their historical collections remained much longer in their more individual but often well-known, rich, reliable and historically-proven local context.

To dig even more deeply into the history of libraries, collections and cataloguing, even the oldest, hand-written, volume-bound, historical catalogues or inventories[13] were always created with a set of rules, even if these may have only developed or may sometimes have been forgotten in compiling the catalogue. They differ in their approaches whether by author/title, subject or shelving, but the core data of author, title, year of publication and, often, place of publication and number of volumes have been present since the earliest catalogues.

These historical catalogues can be considered as the earliest starting point for catalogue records on their way into the bibliographic universe. With the progress and broad user acceptance of electronic catalogues, legacy records and historical collections became more and more integrated into the new information environment. The particular requirements for records to describe these materials exercised a challenging influence on the development of cataloguing codes and data formats. These activities were encouraged by the fear of oblivion, a striving for visibility, technical feasibility, and also the idea of creating a full and comprehensive bibliographic universe. They have not yet been completed and they remain

12 The concept of Universal Bibliographic Control gained importance in the 1960s, cf. Willer and Dunsire (2013, note 11, p. 2 *et seq.*).

13 Their content is very often still of interest for research and has not yet been fully exploited. The care for legacy catalogues must go hand in hand with the care for historical collections.

an ongoing exercise and a challenge. Some projects and their records do much more than describe an existing collection. They allow the reconstruction of lost or dispersed libraries or provide detailed information about unique documents like manuscripts. Today, these catalogues which are to be retroconverted can be considered as valuable primary research data.

Information technology – data formats

With the application of computer and information technology to catalogues, a new decisive standardizing factor was introduced on top of the "traditional" cataloguing code: the data format. If ISBD-areas and punctuation were the first approach to a reliable record structure, the format allowed the next important step towards the modularisation of data. The machine-readable format gives a clear transparent encoded definition of the elements of a record. This implies the possibility of a semantic translation ("naming") of these elements. The data format thus allows the differentiation and naming of information units: for example, the author (as a person or a corporate body); the main title (differentiated from other titles); the year of publication and so on.

Although international standardization coincided from the start with the development of data formats, there was not only one format. Even the use of the same format gave room for individual interpretation, options and enrichment. The continuous development and enrichment of formats in order to better accommodate the diversity of bibliographic records and fulfil the (different and often contradictory) demands of cataloguing rules and cataloguing agencies, and the effort to reconcile these with technological requirements, are characteristic of the development of data formats. Data formats are not a static block like a card catalogue record but are in constant development. They also freed the records and allowed for flexibility within them: additions, deletions, transformations, corrections – these can all be repeated numerous times in one record. The possibility of adding new information may weaken the idea of a "perfect" complete record (dear to those cataloguers who perceive their endeavour as an art form). However, this is not only a loss but a gain for a dynamic future orientation, which even admits the possibility of user input.

Consequent on the introduction and development of data formats, a very significant translation and transformation exercise was carried out on catalogue records, adding machine work to intellectual analysis and practical cataloguing,

and also continuous data maintenance in updating existing records. Internal formats within a cataloguing database system differ from input formats and exchange formats, the first in the care of IT and database specialists, the second in a dialogue between format and rule specialists, data analysts and cataloguers and the third again mainly with data analysts. Format migrations remain on the agenda even after a defined cataloguing project has been finished. Although the application of a format allows for automated data validation, checking the logic of a structure and so on, they are still subject to various forms of human error and are capable of different interpretations. For cataloguers, the knowledge of (at least one) data format should accompany the knowledge of cataloguing rules.

CERL's[14] HPB teaches us that only close co-operation between bibliographically-skilled data analysts and technologically-skilled cataloguers, understanding the format application to their particular cataloguing practices and needs makes for a meaningful and consistent mapping between formats. For the HPB during its initial hosting by the Research Libraries Group (RLG) the central (input, not internal) format was USMARC (then MARC21); UNIMARC was the official data delivery and exchange format. After the migration from OCLC to Göttingen in 2013 the internal format became Pica+; the preferred data delivery format is MARC21. Although excellent political and content-related reasons were advocated, developed and in use for the application of UNIMARC[15] within the HPB, the use of MARC21 (also today for receiving and exporting data) proved that the format in itself only has an auxiliary value to the data. Much more important than any choice of format for a central database is the careful mapping of existing records and their structure to the chosen format. This must take into account not only the structure but must also consider the content (the semantics) stored or delivered in this structure. We have seen in the HPB the extent to which within a format of the same name, the semantics, content, format application or interpretation can differ. Format standardization as carried out over decades and ongoing still is only one, insufficient in itself, approach to this diversity and richness. The harmonization within a single database is an asset before opening up the data again for future use. The situation can be compared to language where syntax (format) and semantics (content) are vital and yet we understand quite a lot of each without fully and correctly applying both rule sets. For the future the

14 CERL is a consortium which stimulates reflections on such issues – the art of cataloguing – within a specialist environment.

15 Mirna Willer was a standing member oft he IFLA Permanent UNIMARC Committee from its establishment in 1991 until 2005, Chair of the Committee from 1997 to 2004, and since then its consultant and an honorary member. Let us remember here the *Application of UNIMARC to Multinational Databases* (1999).

application of central tools to all records in HPB like the checking of place name information against the CERL Thesaurus successfully carried out in 2018 will be as much of an asset for users as a means of contributing enriched data to file providers.

Indexing as a move towards modularization

The introduction and use of a data format allowed for a number of other hitherto unnoticed or presumed choices beyond cataloguing rules. They led to even more individuality and heterogeneity of files in different library systems. In the area of indexing, the filing rules of the card catalogue era and the pre-coordination of search possibilities continue to play a role. How to index which fields has to be determined, in order to allow for string (or phrase) searches and word searches, and how to delimit and index "words". The handling of diacritics, special characters or sequences of letters and numbers, has to be defined. Given the diversity of European languages and scripts, the international character of the records in our catalogues and the different historical layers, it is not surprising that the same issues were tackled in different places by differing, however explicable, decisions. International care for the character set accompanied the development of data formats. The introduction of original script characters and the richness of UNICODE were regarded as a solution. Nevertheless, even today, the same very basic indexing questions arise in every change of library system and in every data transfer and reuse. There is no simple answer to those issues which were not decided within international standardization efforts: for example, the value of the "apostrophe", the hyphen and the umlaut. Again, the HPB as a central database is the best environment for harmonizing these features, allowing reliable retrieval on the basis of the provided data. The new international cataloguing code, *Resource Description and Access (RDA*), first published in 2010, decided to avoid any indexing or filing prescriptions. In a way, this acknowledges and allows the continuation of heterogenous decisions in indexing.

Development of data formats

Data formats have proved to be relatively flexible (the introduction and modification of fields and subfields, indicators etc. is always on the agenda of format working groups) and a number of restrictive issues like a limited num-

ber of characters in a field are questions of the past today. Data format plays a decisive role in the modularisation of records and thus the transition from "records" to "(meta-)data". Although the MARC format still creates a record as an entity,[16] which it defines by a record ID-number attributed by machine, the structure of the fields, subfields, indicators and the thus implied semanticisation of their content, opens the way to new data manipulations. The most evident internationally standardized consequence was the definition of exchange formats like MAB and UNIMARC, with a greater degree of granularity of fields and subfields and their facility for mapping data from different proprietary system formats. This was a huge step towards liberating data out of their silos into the bibliographic universe and allowing us to compare data structures using a common vocabulary. The differentiation of different kinds of data like bibliographic and item specific information was allowed by the format (although already prepared by ISBD). Different services were established taking the records as a starting point. OPACs were individually designed with their own choices about which fields to display in which sequence and with which presentation,[17] their own selection of fields to index and choice of indexing rules, nowadays including the presentation of index entries as facets. They thus allow access to information coordinated by the user, although in the framework of the content taken into account by the OPAC, a fact most users are unaware about.[18] Acquisition and loan systems rely on records and their format as much as today's digitization workflow tools. Thanks to data formats and to building on them, different presentation formats came into being. All of them use a selection of data out of a fuller dataset. Today, it is important to stress this as more and more new formats, seemingly easy and simple to handle, come into existence. From Dublin Core to the emerging IIIF standard, we must be aware that these are derived formats or format applications. They take a selection of data from a much richer and fuller existing structure and present it efficiently according to their individual remit. However, if they were to replace a fully developed bibliographic format like MARC21 or UNIMARC, they would have to reassess and rebuild a universe of features from which, as presentation formats, they can profit without handling and caring for a complete sensibly structured data set.

16 This is also true for the XML-coding preferred for the recording of complex descriptions closer to a dictionary entry than to a catalogue record, as we see in the cataloguing of manuscripts or archival material.

17 Today OPACs have moved far away from the ISBD presentation, which was so useful and internationally shared in the past. Sometimes they hide and omit complex information sacrified to the concept of a "simple approach". As a specialist's tool the HPB aims to present all given data.

18 There is an important awareness issue here for librarians as much as for users. It is an asset of HPB to allow access to the full MARC-record also within the public interface.

A true challenge for bibliographic data formats today is to ensure compatibility with text or collection-focused cataloguing and description formats: from the text-focused TEI-XML-structures to the traditions of museums and archives. Thanks to heritage portals like Europeana giving access to digitized content from different memory institutions, these hitherto closed memory communities can detect identities, similarities and generic differences within their approach to description. This creates an impetus to better share and build common tools which can be efficiently supported by intelligent analysis, understanding and as far as possible mapping of data. Last but not least, the logic framework of RDA has to be fully translated into the bibliographic data format. This is vital for the future of the bibliographic heritage-oriented semantic universe. The work on data format continues: BIBFRAME, initiated in 2011 by the Library of Congress is aimed at integrating the possibilities of linked data and semantic web technologies and at replacing hitherto known bibliographic formats.

Codes, numbers, record types

A well-known feature of data formats is the use of coded information. Codes play an important role in the modularisation of data. Comparable to numeric information, they are much more easily standardized, language independent and are thus easily interoperable. Codes can design semantically important content features, like language or country of publication. They can also design cataloguing or bibliographic structures which are of importance when building full coherent bibliographic information. It makes a difference whether a record describes only a volume in a set (information about the set has to be taken into account as well), an analytical description or a monograph (and even here we would like to know whether further analytical information on the content or for a series is available). Again, CERL's HPB mirrors the diversity of these approaches. Some of the records are not comprehensible taken out of their original context. It also shows that these basic features of the bibliographic universe, multi-volume publications, analytic descriptions, etc. are handled differently everywhere. Even if records are connected within the original related database system through consistent numbering, their transfer into another system demands careful analysis and reconstruction of the internal linking structure. The mere numbers only link in their original context: in a new context they must be considered with particular care in order to continue to ensure this function.

Authority records

A central step in the modularisation of data was the definition and creation of authority records for certain entities, alongside the possibility of linking an authority record to the relevant entity within the data format. In Germany[19] the work on authority records began with corporate bodies. This new entity was introduced with RAK,[20] a national cataloguing code which took into account international developments, including entities thus far unknown or marginalised in German cataloguing. Both the ISBN, and even more so the ISSN, could be considered as early authority numbers, defining a manifestation. In the US, serials and series publications were managed by authority control. Authority records are usually created in a separate file and format environment. Their content, i.e. the information provided about an entity, can be much more developed than within a catalogue record, and is today further enriched by digitized information.[21] Personal names and corporate bodies, titles of works,[22] all forms of subject headings, and in today's terminology, events, places, objects, concepts, have been successfully registered in authority files. The authority format alignment between all these entities as implemented in *Gemeinsame Normdatei* (GND) in 2014, with the introduction of RDA, uses the same modularisation processes as in the bibliographic format. It plays an important role in widening the impact of authority records.

The relationship between authority records and bibliographic records is differently handled in different systems. Relational databases allow for a link between a field in the bibliographic file to a record in the authority file. All too often, the ID-number of the authority record introduced into the bibliographic file, in itself sufficient and best suited for linking, was not considered sufficiently semantically rich. So, the heading of the authority record was introduced in a particular subfield, and the number in another subfield, with consequences for updates in the case of heading changes and transfers into other systems. The CERL Thesaurus, freely accessible on the internet since 2000, containing imprint places, printers, publishers, names of persons and

19 Authority files were created in different ways in different countries. In Germany they were considered as central national files from the start. Again, in sharing at a national and international level, they allowed the opening of systems.

20 *Regeln für die Alphabetische Katalogisierung*, first published in 1976, based on the Paris Principles and ISBD, and developed in a consistent manner was used within German-speaking countries, and was replaced by RDA.

21 For example, portraits of persons, handwriting examples, coordinates for places and the detailed history of text development. Authority work, which was initially meant to provide a controlled access point, developed far beyond this.

22 Just to remember that they are among the earliest IFLA standards: Anonymous Classics, first published in 1964, and 1978.

corporate bodies for early printed books up to the mid-19th century, tried a new approach. Like the HPB, it is also a gathering together of different existing authority or thesaurus files. The Thesaurus is not prescriptive about the heading but allows parallel headings (from different sources). Here the merging of records for the same entity is intended and carried out, although all information coming from different sources is maintained. The CERL Thesaurus was conceived as a search aid to overcome the differences between headings for the same entity in the HPB and later also applied to the CERL portal (accessible since 2005). Today it has also developed as an information system with its own value and is open to other tools like VIAF.[23] It illustrates how valuable this kind of entity-oriented detailed information is; how many different, often research-generated and carefully controlled systems provide such information; and how the information can be usefully gathered together in a central heritage-focused detailed authority environment.

Although the situation is constantly developing and expanding, the authority records have proved their worth as nodes and rich information resources in the bibliographic universe and even beyond. They are important modular entities for sharing information among heritage institutions. Authority files developed by libraries are an asset which museums and archives can reuse, share and build on.

RDF and triples

If this modular approach to a catalogue record as performed by data formats, codes, numbers and authority records was not yet sufficiently visible to be understood and taken into account in a world where interoperability and the opening of hitherto closed data silos was on the agenda, the introduction of RDF (*Resource Description Framework*) in 2014 made clear that well-structured bibliographic information could be divided into a large number of units, and re-organized in so-called triples. A number of semantic relationships are thus made explicit and can be rebuilt in the semantic universe in a huge, growing network. The triples derived from one record, reorganized into numerous individual subject-predicate-object-expressions do not only increase data quantity and value, but they can be interconnected beyond the record, beyond the catalogue and even beyond the bibliographic universe. This does not mean simplifying cataloguing or destroying

23 The feature "same as" gives the same name within VIAF, DBPedia and also a number of contributing authority files.

the coherence of information. On the contrary, it enriches this activity, makes existing or detectable units (entities) and their relationships explicit and more interoperable, and leads them out of the library environment and even beyond the academic world. Here (classical) ISBD could regain its importance as a valuable, standardized way of presenting compact information on a given resource (identified by a record ID). Then this full record can be integrated and connected to the bibliographic, even the semantic universe, by a multiplicity of triples.

RDA and the bibliographic universe

RDA fully builds on modularity and entities and thus acknowledges and furthers the development of the bibliographic universe. It takes into account all resources, even if the core logic (FRBR) was mainly based on manifestations with multiple items. Today, there are still some difficulties in integrating, say, manuscripts, a unique manifestation (or even expression?) with a unique item, a physical object of the real world, which needs to be taken into account as such. But RDA is still in its infancy, and it has a long-term aim: resource description and access. For RDA, modularisation is self-evident, it is entity-based throughout. This goes as far as not taking into account the existing data formats and their necessity of building a complete record, but designing a logical framework. Modularisation calls for decisions on how to apply entities in a model to the material at hand. Again, manuscripts show this. Can we describe every feature at the manifestation level or, for the sake of consistency with imprints and other forms of text transmission, do we want to introduce an item level (for shelfmark, provenance, binding) although both, manifestation and item, are one and unique? When is the content of a manuscript a new work and when is it an expression of a given work? The content of each manuscript might be considered as an individual expression of a work, but this would introduce too much differentiation and incompatibilty with other resources without solving the issue of the object approach (an issue which is shared, for example, with museums). RDA does not give guidance on whether to deal with this type of information in authority records or in the bibliographic format structure, and so this type of advice will be needed for these and numerous other aspects. The modularity of RDA even acknowledges in the four-fold-path all existing traditions and models of recording information: from (un-)structured description to (international URI) identifiers. Thus a rule set or, better, an application profile

"RDA and manuscripts" must not limit itself to the definition of a (short) record translated into a bibliographic format. It can include descriptions and long text passages often coded in manuscript databases with TEI or EAD structures and living outside the classical bibliographic database context.

Modularisation frees the records, opens the catalogue, and allows for interoperability in the semantic universe. The future will bring about interlinked networked information where the user will be able to contextualize according to individual needs. For this the nodes (modules, entities) are the decisive features. They need to be defined (RDA proposes a framework for this) and interconnected or related. Authority records, codes and standard identifiers will play a prominent role. The more they are harmonized and standardized at an international level, the better. In addition to ISBN and ISSN, there is already a rich portfolio of ISO-certified identifiers, but how well are they implemented and even defined in the light of RDA? In the area of manuscripts, a new identifier has been advocated since 2017: an ISMI (International Standard Manuscript Identifier). It designates the manuscript as a real physical object, as traditionally does its shelfmark, which is however not easily interoperable.[24] This initiative joins the interests of a group advocating for a cultural heritage entity (CHE) identifier.

Although we recognize the emergence of this world of interoperability and semantic networking, we cannot yet see it really working. On the contrary, we can experience how difficult, even in the context of a firmly backboned technical infrastructure, the building of a database like the HPB, which is fed by various files, still is. The alternative relying on central, consistent indexing of remote files (as is the case in the CERL portal) needs the same careful analysis and the same structure and content and context-oriented mapping. A central database may be "a Procrustean bed", but it also promises transparency and a form of reliability: and it allows for a quality-controlled presentation of content and structure elements. The future of small individual databases is much more threatened, as we see by a number of activities integrating them into larger contexts, like the HPB, the necessity of their constant migration or today their adaptation to RDA standards. Last but not least, the requests for hosting individual specialist databases are growing, as is shown by the CERL initiative started in 2016. This does not only mean long-term archiving of data but keeping alive useful information in a sensible presentation.

24 For example, it needs to be specified by naming the collection, using the International Standard Identifier for Libraries, ISIL.

Databases as much as (meta-)data, as much as the books on our shelves, need continuous curation. We need data scientists and data managers; we need not only cataloguers but cataloguers who are data artists: we need them at an international level as much as at a national, regional and even local level for large libraries and information systems. Their work is not only about analysis; it is about building a synthesis, forming a network, making data interoperable. Sharing in this continuous endeavour in an international environment is both hard work and a privilege, and it is an important service to the future of heritage. Thank you, Mirna, for the admirable and vital part you have personally played.

References

Application of UNIMARC to Multinational Databases. (1999). Feasibility study coordinated by C. Fabian; Report compiled by A.G. Curwen and C. Kirk. München: De Gruyter Saur.

CERL portal, http://cerl.epc.ub.uu.se/sportal/ (05-02-2019).

CERL Provenance Digital, http://arkyves.org/r/cerl/pda (05-02-2019).

CERL Thesaurus, https://data.cerl.org/thesaurus/ (05-02-2019).

Consortium of European Research Libraries (CERL), https://www.cerl.org/ (05-02-2019).

Barbier, F., Dubois, T. and Sordet, Y. (2015). *De l'argile au nuage – une archéologie des catalogues: (IIe millénaire av. J.-C. – xxie siècle)*. Paris: Editions Des Cendres, http://www.lescendres.com/livre/de-largile-au-nuage#! (05-02-2019).

Gemeinsame Normdatei (GND), https://de.wikipedia.org/wiki/Gemeinsame_Normdatei (05-02-2019).

Regeln für die Alphabetische Katalogisierung (RAK), https://de.wikipedia.org/wiki/Regeln_f%C3%BCr_die_alphabetische_Katalogisierung (05-02-2019).

Répertoire international des sources musicales (RISM-OPAC), http://www.rism.info/home.html (05-02-2019).

Resource Description and Access (RDA), https://en.wikipedia.org/wiki/Resource_Description_and_Access (05-02-2019).

Universal Short Title Catalogue (USTC), https://www.ustc.ac.uk/ (05-02-2019).

VIAF: The Virtual International Authority File, https://viaf.org/ (05-02-2019).

Willer, M. and Dunsire, G. (2013). *Bibliographic Information Organization in the Semantic Web*, Oxford: Chandos.

STRUKTURA I SEMANTIKA, KOHERENCIJA I MREŽE – ŽIVI BIBLIOGRAFSKI SVIJET

RAZMIŠLJANJA LJUBITELJICE KATALOGA U ČAST PODATKOVNE ZNANSTVENICE MIRNE WILLER

Ključne riječi:
europska tiskana knjiga (1450. – 1830.), Heritage of the Printed Book (HPB), upravljanje podacima, podatkovna znanost

Sažetak

Sjećajući se zajedničkog rada u Konzorciju europskih znanstvenih knjižnica, a osobito u stvaranju HPB-a (danas Heritage of the Printed Book Database), ova razmišljanja o „živom bibliografskom svijetu" prate razvoj kataloga i strukture zapisa od monolitnog pristupa prema modularnoj, semantički bogatoj mreži. Važnost upravljanja podacima i doprinos podatkovne znanosti velikoj viziji bibliografskog svijeta ilustrirana je sa stajališta nekoga tko nije stručnjak, ali je ljubitelj kataloga, i to na način koji je lako razumljiv svima koji se susreću s takvim pitanjima. HPB, baza podataka koja daje pristup starijoj europskoj tiskanoj knjizi (1450. – 1830.), ovdje se smatra „mikrokozmosom" koji omogućava uvid u opća pitanja bibliografskog svijeta.