Adapting CERIF for a national CRIS : a case study

Kremenjaš, Davorin; Udovičić, Petra; Orel, Ognjen

Adapting CERIF for a national CRIS: A case study

D. Kremenjaš*, P. Udovičić * and O. Orel*
* University of Zagreb, University Computing Centre, Zagreb, Croatia
{davorin.kremenjas, petra.udovicic, ognjen.orel}@srce.hr

***Abstract* - A development and implementation of a national Current Research Information System (CRIS) in Croatia is on its way. During the design of the system, global trends in the area of research information management and exchange were considered. The initiative to share knowledge among a broad variety of stakeholders in the national research community resulted, among other, with CERIF case study.**

CERIF is a research information data model recommended by the EU and governed by euroCRIS. From a broad point of view, it meets the initial requirements for the system.

In the area of semantics and multilingualism CERIF is effective and valuable. Nevertheless, at the very beginning of implementation, functionality gaps emerged and it was necessary to make certain adjustments, especially regarding temporal aspects. With the intention to achieve the resolution of observed gaps within the model and to satisfy key stakeholders' concerns, architectural modifications and extensions of the model were made.

This paper includes the presentation of Croatian CRIS, CERIF and the case study explaining what is needed to adapt CERIF for a national research information management system and vice versa, all this while paying attention to the critical aspects.

Keywords* - *CRIS; research metadata; CERIF; information systems; temporal databases

I. INTRODUCTION

Research information management (RIM) is a field of work whose primary areas of concern are aggregation, curation and utilization of data regarding research activities [1]. In order to effectively perform these tasks, a new class of information systems emerged in the last decade, often referred to as Current Research Information Systems (CRISes) [2]. These systems are "current" in the way they hold information which is actual and will enable analyses that are presently important, in addition to all other historic information.

In this paper, we show varieties of CRISes and current state of affairs regarding research information in Croatia, including a case for creating a national CRIS. We will present some of the common data models used for RIM and consider one of them, Common European Research Information Format (CERIF) [3] in detail. A substantial amount of work deals with informational aspects of CERIF (e.g. [4,5,6]). Main contribution of our work is the implementation of CERIF in a huge research information system, while taking all the technical aspects into account. Out of several data models, CERIF was selected as a base for this system mainly because of its completeness, institutional support and flexibility.

We discuss the abilities and limitations of common relational databases to support such a model and consider changes in CERIF in order to achieve a level of practicality and ease of use for both system architects and developers.

This paper is organized as follows. In section II CRISes are discussed in more detail and a case for a national CRIS implementation in Croatia is presented. Section III shows common research information data models used today and gives more information about CERIF. In section IV challenges of implementing CERIF in a relational database are presented, along with some adaptations in order to make it more suitable for common relational databases, while the last chapter concludes the paper.

II. A NATIONAL CRIS IN CROATIA

The scope of CRISes varies. Some only contain information about scientific publications or projects, while some contain full-scale information about researches, institutions, publications, journals, projects, patents, events (such as conferences or congresses) and many more. Additionally, most CRISes are localized to a specific institution or a university, while the cases for a national CRIS are rare [7]. Some CRISes are implemented at a national level, as aggregators of information already present in institutional CRISes, like the ones in Finland [8] and the Netherlands [9]. There are also examples of a single-installation national CRISes, where all users directly connect to and use a central system, like the ones in Norway [10] and Slovenia [11].

A. Research Information in Croatia

The project Scientific and Technological Foresight (STF) is underway in Croatia. This project has several goals, one of which is creation of the national CRIS. In the project preparation, a political decision was made to create a new system instead of buying an existing one, and possibly adjusting it to specific Croatian needs. Currently, there are various "islands of informatization" regarding research information in Croatia. There are some systems which cover only a part of all information of the entire future CRIS - publications, equipment, projects, researchers, institutions [12,13,14]. These databases are

limited in scope and barely interconnected. On top of that, none of the CRISes should be isolated, so it is expected that this system is interoperable to other national and international systems and platforms (OpenAIRE [15], ORCID [16], etc.). Even though CRIS is not a data repository, it is a good practice to interlink CRIS with institutional repositories that usually hold full texts of the publications, research data etc.

B. Motivation for a Single-Installation CRIS

When considering the possible architectures of the national CRIS, the main dilemma is whether to create an aggregator of all the information available in existing systems and use it for analytic purposes, or to create a single-installation system, which will be used by all end-users. While the first option means that existing systems will continue to be used and most of all users should not notice any change, it also means additional applications should be made in order to cover the information which is currently not collected in any way (e.g. events, services, ...) and a huge part of interoperability features would need to be implemented in all those systems and applications. In this case, creating a single-installation national CRIS which will replace some existing systems and introduce many new features makes more sense, for following reasons:

- having a single source of truth for all the research information in the country
- simplicity of not dealing with interoperability between many smaller applications
- total cost of ownership (including maintenance, further features, etc.).

Croatian CRIS (CroRIS) will therefore be created and implemented as a single-installation system. At the beginning of the project, a planning phase was conducted, and a data model was chosen and the system itself is currently being developed. The whole system will be developed and implemented into production in stages, starting with official researchers and institutions registers, followed by the first version of CroRIS portal, etc. During each phase of implementation, a new set of users will be introduced to the system, along with import of relevant data from other systems and API subsystems, to provide the interoperability platform for other national systems. It is planned that the initial development and implementation will take more than two years.

III. Other Data Models for Research Information

As stated earlier, CERIF was selected as the basic data model for CroRIS. In this section, we give a short overview of some of the other actual data models for research information and state main shortcomings that prevented their further consideration.

A. Component MetaData Infrastructure

Component MetaData Infrastructure is a part of larger CLARIN (Common Language Resources and Technology Infrastructure) project which aims to create a research infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular the humanities and social sciences [17].

CMDI is CLARIN project's framework for metadata reuse [18], using components and profiles as containers and main building blocks to be used later for operational systems implementation. It is a very high-level/conceptual framework similar to Dublin Core Metadata Initiative, currently having only early-stage prototypes with limited functionality.

B. Data Documentation Initiative

Data Documentation Initiative (DDI) [19] is an international, free XML based standard backed by high profile institutions such as MIT, UCLA and Eurostat [20]. Its main target is a limited set of sciences: social, behavioral, economic, and health. It has strong support for controlled vocabularies. Special care has been taken in the area of mapping to existing standards such as Dublin Core and MARC. It boasts a state-of-the-art documentation, but it has no direct mapping to relational data model.

C. B2FIND

B2FIND is a discovery service for harvesting existing/external repositories [21] rather than a data model to implement own research information repository/CRIS. It is supported by another well-established European research infrastructure provider entity – EUDAT – and their Collaborative Data Infrastructure [22] EU wide project (focusing on storage and network infrastructure specialized for data hosting) to support research activity. It has limited support for data modeling.

IV. CERIF in Detail

A. Origin and Basics

CERIF dates back to 1970s but only in 1987 was the first version developed and formally released [23]. Since its beginnings the model improved over time. Significant development of the model started with CERIF 1991. The model included only research projects, persons, organizations, publications, equipment, facilities and a few other entities. A need for a more contextual information was recognized later on, together with the functional dependencies problem.

First formally developed and released version of CERIF was developed by Research Database working group organized by EU. During the year 2000 custodianship of CERIF was handed over to euroCRIS, with the aim to achieve interchange, integration and standardization of European Research Area.

CERIF 2000 standard showed significant improvement in the area of describing the research domain. Main entities were interconnected through new attributes: role and date/time. Other mentionable improvements that came with this version were the implementation of multilingualism, storing contact information's in one entity and inclusion of a single classification system or controlled vocabulary.

More flexibility with capturing the semantics of relationships was achieved with the release of CERIF 2008 [24]. The model form as recognized today was built as a result of dividing entities into five groups:

- Core CERIF entities – Project, Person, OrganisationUnit, ResultPublication.
- 2nd Level CERIF entities – introduced with intention to capture the context of activity and interaction. Some of those are Service, Event, Facility, Citation...
- CERIF Relationships (LINK Entities) – those entities with the structure like Entity1Name_Entity2Name with goal to describe relationships among entities while considering temporal aspect.
- Language-dependent CERIF Entities – entities requiring representation in more than one language (Title, Abstract, ResearchInterest...)
- CERIF Semantic Layer – generically made with the intention to allow representation of virtually any kind of relationship within the model.

Later releases of CERIF made improvement on entities such as research infrastructure entities (Facility, Equipment, Service), together with the geographic bounding box in the same context. New entities were created (Medium entity), new attributes were added and semantic formalities were extended (vocabulary).

Attribute values in language-dependent entities of CERIF can be stored in any language. Values can be machine and/or human translated. Relations among entities are represented with a time range of validity (start date and end date for the relation).

CERIF used today contains 295 entities, 1829 attributes, 295 primary and 3 alternate keys as well as 638 relationships.

B. Semantics and Vocabularies

Standard relational data models often employ a relatively large number of domain-specific *type entities* which correspond with certain parts of controlled system vocabularies and can be visually represented as *leaves* in ER (entity-relationship) model, to be later referenced by fact-oriented data entities through referential integrity rules. Depending on the size of the overall relational model the number of such type entities can range from just a few to hundreds in larger systems.

A distinct feature of CERIF model dealing with such type entity proliferation is its *semantic layer* which minimizes the number of such hierarchies and employing a meta-layer for modeling types, categories, statuses, flags, varieties and similar concepts instead. Such concepts in CERIF are mapped to *classifications* which are further grouped into *classification schemes*. Each classification and classification scheme have several attributes helping to describe them. Any further type-based modeling becomes unnecessary as all main data model entities now reference this new meta-model comprised of just a handful of tables (as in an example of organizational structure depicted in Figure 1.) regardless of the size of the overall data model. What was previously modeled through a scheme structure now follows semantic rather than structural hierarchy. Any links between main entities also become classified through the above mechanism, allowing for virtually endless possibilities for expressing semantic relations between different entities.

This further enables easier handling and a lot more control over formal vocabularies used both within the system, but particularly for external systems data exchange where each term must be recognized and agreed upon between all the communicating parties. If deemed necessary at some point it could also become a key enabler of *semantic web*, a much-discussed concept for over a decade now but with very slow adoption worldwide.

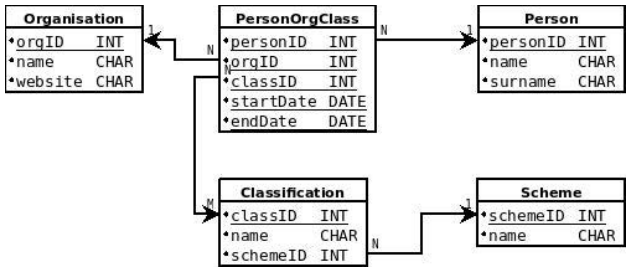

Figure 1. An organizational structure represented using Classification/Scheme

C. Why CERIF?

Among a multitude of research information targeted data models, CERIF stood out as the most appropriate for CroRIS project for a number of reasons:

- "open source": not controlled by any commercial entity; managed by euroCRIS
- strong institutional support: backed by the EU Commission
- standards based: SQL + XML
- simple: in comparison to more abstract/rich generic semantic frameworks such as ones listed above
- designed from ground up as a relational data model - our area of expertise
- extremely flexible: capable of modeling and meta-modeling just about any high-level concept using its semantic layer
- strong support for vocabulary formalization: key enabling point for EU-wide (potentially global) CRIS systems standardization and interoperability
- easy integration with external systems: OpenAIRE as a minimum, but a lot more to come
- adopted: the number of institutional, regional and national CRISes in the Directory of Research Information Systems (DRIS) [25] which claim to be CERIF compatible or use the software which is marked as CERIF compatible is 267.

Downsides:

- incomplete/inappropriate temporal aspect
- sporadically developed - several years of pause between development projects (current CERIF refactoring project started in January 2020 [26])
- no reference implementation/blueprint - some relational DB schema and XML docs but it is outdated.

Given these pros and cons of CERIF, it should be noted that, for CRIS purposes, CERIF is the most complete data model and the most suitable one.

V. CERIF Challenges and Adaptations

Relational databases still make the core of most of the big systems, in the face of recent popularity of other types of databases, namely NoSQL/NewSQL. For many reasons, a relational database will be a core of CroRIS as well.

Although CERIF provides an excellent starting point, it requires many changes and adaptations to be used efficiently and to cover all functional requirements. The below list is not exhaustive, but it still depicts a sheer number of changes introduced during the initial system design phase where a minority of planned features were covered.

A. Temporal Aspect

In CERIF, a special attention is given to the temporal information aspect. When it comes to temporal relational databases, CERIF as a data model focuses on valid time [27]. Therefore, valid time support in current relational database management systems (RDBMS) is of a special interest in our work.

Temporal support in relational databases is defined in the SQL:2011 standard [28], but current RDBMSs are yet to implement it in this regard. Many of RDBM's vendors support only system-time aspects. Implementing valid time in RDBMS means the database must support a "period" data type (a fixed interval in time), operators to deal with it [29] and temporal primary and foreign key constraints. Of these three, the most difficult is to support temporal primary and foreign keys because, in a typical relational database, these constrains are always checked only for equality, while in temporal regard the temporal part of keys must be checked using interval operators.

The most progress in implementing application-time temporal features has been made in PostgreSQL [30] and IBM DB2 [31]. Currently, PostgreSQL still lacks temporal foreign keys feature, while DB2 has temporal foreign keys included in the last version, but still with some minor restrictions.

CERIF strives to fully incorporate temporal data aspects from ground up, but falls very short of its intention, an area which proved to be the biggest obstacle to a real-world implementation in our case and probably an area of our biggest deviation from the proposed model. An example of this is presented in Figures 2a through 2c where a case of a subset of entity attributes are known to have their own temporal characteristics, but with no appropriate support in current CERIF to support it. A schema evolution is shown starting with Figure 2a. where only a Country entity has temporal aspect (supported by the current CERIF schema) through Figure 2c. where a country name has its own multilingual (supported by CERIF), but also independent temporal aspect (not supported).

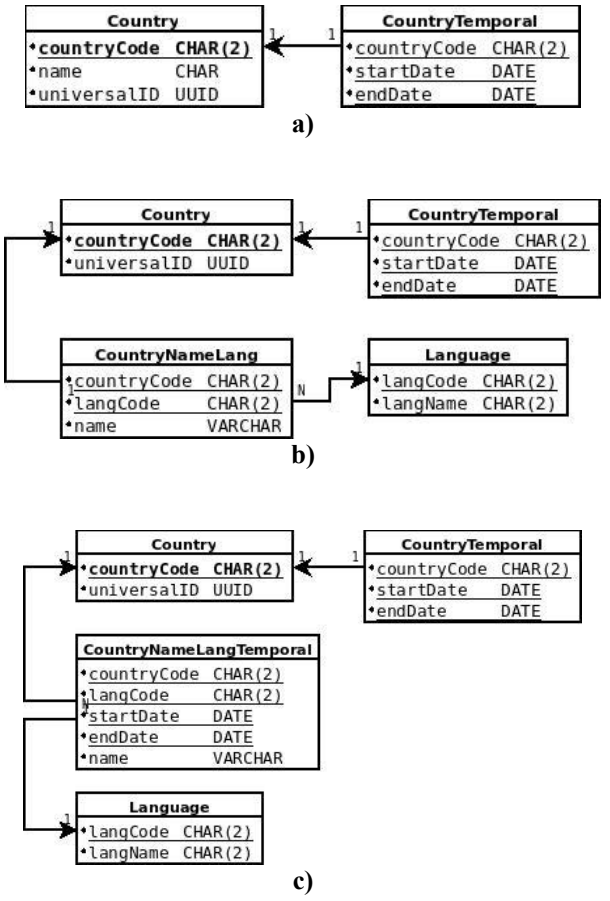

Figure 2. Multi Language Temporal Attribute within Temporal Entity

B. Changing UUIDs from CHAR (128) to CHAR (36)

A portion of the semantic layer is comprised of globally unique identifiers (UUID). An UUID is 128 bits long and can guarantee uniqueness across space and time [32]. Although well-known and easily implemented from a technical point of view, they pose a serious challenge from a data exchange perspective where disparate systems need to be fully in sync with their formal vocabularies and UUIDs representing terms of such vocabularies. As there is no European and especially no global standard to be followed in this respect within the area of research information, the only sensible approach was the adoption of the common denominator strategy, which in this case resulted in using somewhat limited OpenAIRE vocabularies.

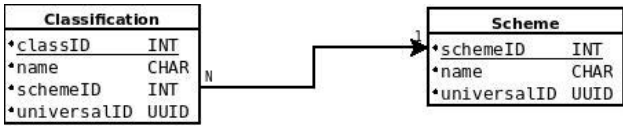

Figure 3. Classification/Scheme UUIDs

In CERIF it has been generically defined as CHAR(128) and assigned to all main entities (Figure 3.) presumably to store all bits as ones and zeros within a CHAR/string, but it is more efficient to use either a binary data type available from the underlying DBMS or a standard string-based representation which is 36

characters long (32 alphanumeric characters and 4 hyphens), e.g. 932f3658-d98c-21e2-b365-515746359182.

C. *Removed many NOT NULL constraints*

Numerous attributes within the existing CERIF model are defined as non-nullable, but for many of them there are no values defined in any publicly available standard. For example, sources for many classification definitions originally envisaged as being a part of CERIF controlled vocabularies are not available. As a consequence, all such attributes have been redefined as nullable.

D. *Descriptive attributes expanded from CHAR (127) to LVARCHAR*

A number of object definitions and descriptions did not fit into predefined templates and had to be expanded to 2k or even up to 16k variable character types.

E. *Switched from CHAR to VARCHAR for most attributes*

VARCHAR semantics proved to be a lot more flexible within the database layer and also more appropriate for object-oriented programming.

F. *Removed/Ignored unary *_Class tables where deemed unnecessary*

Most CERIF entities have a corresponding *_Class table (Figure 4.).

Our understanding of such unary classification tables is a re-affirmation of main object's existence plus their further classification and temporal limitations. In most cases such extra information was not necessary (in contrast to binary relationships where similar tables are always needed) and was removed.

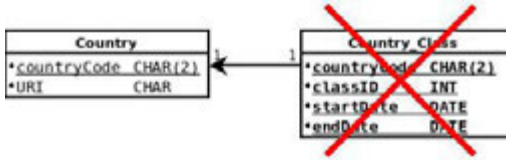

Figure 4. Country and its Classification Table

G. *Wrapped SELECT operations with views*

Due to underlying data model complexities as described above application read access to table data has been simplified and standardized using database views where possible.

H. *Wrapped INSERT/UPDATE/DELETE operations with stored procedures*

Data write access proved to be even more challenging as inserts/updates/deletes of one logical construct seen from the application point of view most often mapped to several database level objects. In this case database stored procedures proved to be of use, even though it is not the preferred approach in object-oriented programming world. For us they clearly demarcated responsibilities in each software code layer and provided a much simpler programming interface for application developers.

I. *Removed classification scheme ID from all primary and foreign key definitions*

Composite primary and foreign keys in the current CERIF version (1.6) proved to be unnecessary, adding complexity but not functionality and were removed (used simple keys instead). This is also a change scheduled for the next official CERIF version (1.7/2.0?).

J. *Removed UUID data type*

The research showed there is no standard, controlled vocabulary defining UUIDs. They were removed and primary and foreign keys were redefined as integers. This decision was taken with understanding once a widely adopted standard emerges, a key mapping to interface with the outside world will have to be defined.

K. *National Registers*

National specifics when dealing with scientific and research sources of information also had to be taken into account. Where CERIF did not provide an adequate data model to cater for local science and research information landscape, the necessary tables were added, modeling them to follow CERIF guidelines with a different table name prefix to be easily distinguishable.

L. *Added additional attributes into existing CERIF-defined tables*

A variation on the above comprised of expanding existing CERIF tables with new attributes where not all required information was already covered.

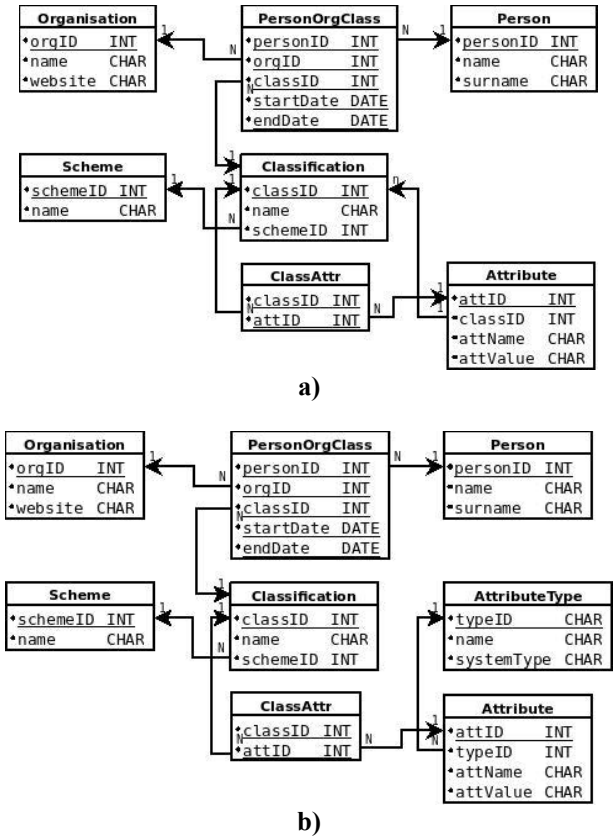

Figure 5. Organizational structure: a) no semantic layer for attribute type; b) using semantic layer for attribute types

M. Attribute Meta-modeling

Even the powerful semantic layer meta-modeling sometimes falls short when faced with more complex cases of abstract concepts. Non-trivial object classifications which themselves have extended attributes (thus evolving into first-order objects rather than just simple object classifications) frequently occur during modeling new or extending existing entities. Examples in Figure 5a. and 5b. solve such cases in a similar way: the latter using a "standard" approach; the former using CERIF-like semantic approach. This type of problems remains to be solved in our current model even though there is a general recognition of its existence and potential solutions for successful incorporation in the overall data scheme.

VI. Conclusion

In this paper we presented information systems focused on management and analyses of research information – CRISes. From our perspective of building a national CRIS, we presented some potential data models that could be used in such a system, emphasizing CERIF as the most developed one. Unlike the most of the previous work regarding CERIF, we focused on it from the technical point of view. As the main contribution of our work, we presented both positive and the negative sides of implementing it in the relational database and suggested a series of adaptations in order to make it more suitable, especially when implementing CRIS on a national level.

Our future work will include the continuation of CERIF adaptation regarding possibilities of relational databases and the needs of various stakeholders. We will focus on maintaining the most important ideas and concepts of CERIF, while at the same time enhancing it with new data in the national and international context, and giving a special attention to ease of use of temporal data and semantic layer.

References

[1] Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Hoggard, S., Mercer, H., ... & Wright, J. (2017). Research information management: defining RIM and the library's role. OCLC Research.

[2] euroCRIS: Why does one need a CRIS? (https://www.eurocris.org/why-does-one-need-cris) [9.1.2020.]

[3] euroCRIS: Main features of CERIF (https://www.eurocris.org/cerif/main-features-cerif) [9.1.2020.]

[4] Jeffery, K., Houssos, N., Jörg, B., & Asserson, A. (2014). Research information management: the CERIF approach. International Journal of Metadata, Semantics and Ontologies, 1, 5-14.

[5] Ivanović, D., Surla, D., & Racković, M. (2011). A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics, 86(1), 155-172.

[6] Macan, B. (2015). Model sustava informacija o znanstvenoj djelatnosti za hrvatsku akademsku zajednicu (Doctoral dissertation).

[7] euroCRIS: Directory of Research Information Systems (DRIS) (https://www.eurocris.org/eurocris-directory-research-information-systems-dris) [9.1.2020.]

[8] Puuska, H. M., Guns, R., Pölönen, J., Sivertsen, G., Mañana-Rodríguez, J., & Engels, T. (2018). Proof of concept of a European database for social sciences and humanities publications: description of the VIRTA-ENRESSH pilot.

[9] Jippes, A., Steinhoff, W., & Dijk, E. (2010, December). NARCIS: research information services on a national scale. In International Conference on Open Repositories: Proceedings.

[10] Karlstrøm, N. (2016). More than just a CRIS-How CRIStin is building a national infrastructure for Open Access.

[11] Korosec, A. (2014). SICRIS, V3. Organizacija znanja, 19(1), 22.

[12] Croatian Scientific Bibliography CROSBI (https://www.bib.irb.hr/) [9.1.2020]

[13] Šestar equipment database (https://sestar.irb.hr/en/) [9.1.2020]

[14] Croatian Science Foundation Projects database (https://www.hrzz.hr/default.aspx?id=1205) [9.1.2020]

[15] OpenAIRE (https://www.openaire.eu/) [9.1.2020]

[16] ORCID (https://orcid.org/) [9.1.2020]

[17] Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In 6th International Conference on Language Resources and Evaluation (LREC 2008).

[18] Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In LREC 2012: 8th International Conference on Language Resources and Evaluation (pp. 1387-1390). European Language Resources Association (ELRA).

[19] Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. International Journal of Digital Curation, 3(1).

[20] Gregory, A. (2011). The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes. Open Data Foundation.

[21] B2FIND (https://www.eudat.eu/services/b2find) [9.1.2020]

[22] Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R., Coveney, P., & Baxter, R. (2013). EUDAT: a new cross-disciplinary data infrastructure for science. International Journal of Digital Curation, 8(1), 279-287.

[23] Asserson, Anne, Keith G. Jeffery, and Andrei Lopatenko. "CERIF: past, present and future: an overview." (2002).

[24] Jörg, Brigitte. "CERIF: Common European Research Information Format-Insight into the CERIF 2008-1.1 Release." In CRIS. 2010.

[25] Directory of Research Information System (DRIS) (https://dspacecris.eurocris.org/cris/explore/dris) [26.5.2020]

[26] CERIF refactoring project (https://www.eurocris.org/cerif-refactoring-project-introduction) [26.5.2020]

[27] Jensen C.S., Snodgrass R.T. (2009) Valid Time. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

[28] Kulkarni, K., & Michels, J. E. (2012). Temporal features in SQL: 2011. ACM Sigmod Record, 41(3), 34-43.

[29] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11), 832-843.

[30] PostgreSQL (https://www.postgresql.org/) [9.1.2020]

[31] IBM DB2 (https://www.ibm.com/analytics/db2) [9.1.2020]

[32] A Universally Unique IDentifier (UUID) URN Namespace (https://tools.ietf.org/html/rfc4122) [9.1.2020]