Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

EPOS: a novel use of CERIF for data-intensive science

Daniele Bailo[a*], Keith G. Jeffery[b]

[a] *INGV - Istituto Nazionale Geofisica e Vulcanologia, via di Vigna Murata 605, 00143, Rome, Italy (daniele.bailo@ingv.it)*
[b] *Keith Jeffery Consultant, Shrivenham, United Kingdom (keith.jeffery@keithgjefferyconsultants.co.uk)*

Abstract

One of the key aspects of the approaching data-intensive science era is integration of data through interoperability of systems providing data products or visualization and processing services. Far from being simple, interoperability requires robust and scalable e-infrastructures capable of supporting it. In this work we present the case of EPOS, a plan for data integration in the field of Earth Sciences. We describe the design of its e-infrastructure and show its main characteristics. One of the main elements enabling the system to integrate data, data products and services is the metadata catalogue based on the CERIF metadata model. Such a model, modified to fit into the general e-infrastructure design, is part of a three-layer metadata architecture. CERIF guarantees a robust handling of metadata, which is in this case the key to the interoperability and to one of the feature of the EPOS system: the possibility of carrying on data intensive science orchestrating the distributed resources made available by EPOS data providers and stakeholders.

Peer-review under responsibility of euroCRIS.

Keywords: Research Infrastructure, e-infrastructure, data integration, data intensive, metadata, cerif, epos

1. Introduction

In the last few decades the way of carrying on research has changed radically. According to some authors[1] research passed through different eras since the beginning of science, and in the present days we are leaving the computational science era and moving to a new chapter of science history: the data-intensive science era, where the

* Corresponding author. Tel.: +39 06-51860728 Fax. +39 06-51860507
E-mail address: daniele.bailo@ingv.it

Peer-review under responsibility of euroCRIS.

amount of produced data outstrips our capacity of collecting and analysing it and where "the goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other"[2]. Interoperability is therefore fundamental. Another complementary view, carried on by other authors[3] states that integration of data, which is one of the main reason for system interoperability, is another core concept: "Data integration is crucial […] for progress in large-scale scientific projects, where data sets are being produced independently by multiple researchers [and] for better cooperation among government agencies".

Since interoperability and data integration are so important, it is possible to observe that theoretical models and e-Infrastructures design papers for system interoperability and for data integration, are plentiful in scientific literature: we span from theoretical models developed by huge companies and academic researchers[3,5],to database models and implementation of seismogenic sources[4], to reviews and summaries on the topic of integration[18] and many others. Together with such diverse, but often complementary, scientific and technical visions, there is an outstanding number of local e-Infrastructure implementations, of local data formats, of metadata standards and data delivery systems which only in a few cases are operative and able to share data with common standards, as in the fortunate case of some Earth sciences like seismology[6] .

Given this highly scattered scenario, the fundamental but very demanding task of integration is usually carried on by European-wide organizations or European projects, which can afford long term vision both in term of human and economic resources.

This is exactly what happens in the field of astronomy, astrophysics and remote sensing, where huge organizations such as European Space Agency (ESA), National Aeronautics and Space Administration (NASA), Japan Aerospace Exploration Agency (JAXA) and many others can manage and coordinate in a consistent way all the resources required to carry on research: satellites, telescopes, other sensors or machinery and also the e-Infrastructure which enable researchers to retrieve, store, exchange and elaborate data, thus reducing the amount of different data and metadata formats, software and procedures to elaborate data etc.. When just one organization or a small group manages all the resources required to carry on research in a certain science field, then there are good chances that an e-infrastructure with a high interoperability factor is designed and implemented.

Unfortunately this is not the case of Earth Sciences, where the relatively low cost of sensors and hardware (storage, servers, seismic station, GPS station, gravimetric stations, magnetometers etc.) required to create a research infrastructure (RI) allow any institution to create its own implementation of the RI and of the underlying e-infrastructure. To overcome this, projects devoted to create synergies among different disciplines have been created, as EUDAT[13] , which deal with the creation of a e-layer to store, securely preserve and cure the data and discovering it or Global Earth Observation System of Systems (GEOSS[14]), devoted to "proactively link together existing and planned observing systems around the world and support the development of new systems where gaps currently exist". In both cases the common goal is to provide a certain degree of integration of data and datasets coming from different science fields, so that a user might be able to discover geological and linguistic data information with one discovery action (query). Although this global action is important and fundamental, dealing with such diverse data may pose intrinsic limitations to the discovery process.

A peculiar case, studied in this paper, is the European Plate Observing System (EPOS) which deals with data coming from solid Earth Sciences In this case, both community building actions, necessary because of the community high scattering level, and design of e-infrastructure are very ambitious because they are confined in the boundaries of solid Earth Sciences. This project, whose aim is to have a real integration of science data and common access to services from one single integrated online environment, rely on the constructions of the EPOS Integrated Services. They utilise a metadata engine which is one of the main components of the system, and which relies on a specialized, extended implementation of the Common European Research Infrastructure Format (CERIF) model expressly tuned to carry on data intensive science.

In this paper we describe the main features of such system, metadata catalogue and its implementation.

2. European Plate Observing System

EPOS mission is to integrate existing, but also new, distributed research infrastructures (RIs) for solid Earth Sciences warranting increased accessibility and usability of multidisciplinary data from monitoring networks,

laboratory experiments and computational simulations. This is expected to enhance worldwide interoperability in the Earth Sciences and establish a leading, integrated European infrastructure offering services to researchers and other stakeholders.

Indeed, progress in the understanding of the physical processes controlling earthquakes, volcanic eruptions, unrest episodes and tsunamis as well as those driving tectonics and Earth surface dynamics, requires a long-term plan to facilitate integrated use of data, models and facilities from existing, but also from new, distributed research infrastructures for solid Earth Sciences. EPOS represents such a plan and expresses a scientific vision and an IT approach in which innovative multidisciplinary research is made possible for tackling this challenge.

Some of the keywords often mentioned when referring to EPOS are: multidisciplinary, integrated, holistic, comprehensive, efficient, e-Science, all clearly identifying EPOS as a tool to make integrated use of data, data products, software and services (including laboratories) provided by different research infrastructures operating in the solid Earth Sciences domain. Hence, EPOS is not only a portal to domain-specific (thematic) datasets for download: EPOS is this, and much more.

The ambition of EPOS is to overcome the general complexity a researcher must face when using a wide diversity of data and data products to make her/his research, by providing a simple, “one-stop shop” tool. In short: EPOS is more than data download, data mining, data discovery.

The technical goal is to provide an integrated environment where the user can browse, preview and/or select data, and then simply download them or perform processing and modelling directly online.

To create such an integrated environment an e-Infrastructure must be designed.

3. E-Infrastructure design

In order to organize and properly manage all the intraction between different EPOS actors, EPOS architecture is structured as follows (Fig.1):

- *Integrated Core Services (ICS)* provide access to multidisciplinary data, data products, and synthetic data from simulations, processing and visualization tools. However, because EPOS means to integrate, analyse, compare, interpret, and present data and information, ICS does not simply mean data access, they are the place where integration occurs
- *Thematic Core Services (TCS)* are infrastructures that provide data services to specific communities (they can also be international organizations, such as ORFEUS[17] for seismology).
- *National Research Infrastructures and facilities* provide services at national level and send data to the European thematic data infrastructures.

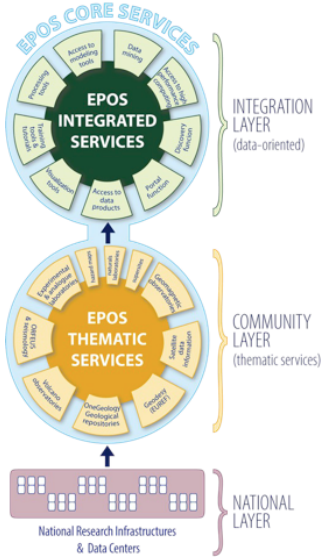

Figure 1: EPOS architecture. The diagram shows the three layers in which the EPOS components (institutions and services) have been organized: National Layer, Community Layer, Integration Layer.

The main concept is that the EPOS TCS data and services are provided at the layer where the integration occurs, that is to say the ICS. It will happen by means of a communication layer called the *compatibility layer*, as shown in the functional architecture (Fig. 2). This layer contains all the technology to integrate data, data products and services from many communities into a single integrated environment: the Integrated Core Services (ICS).

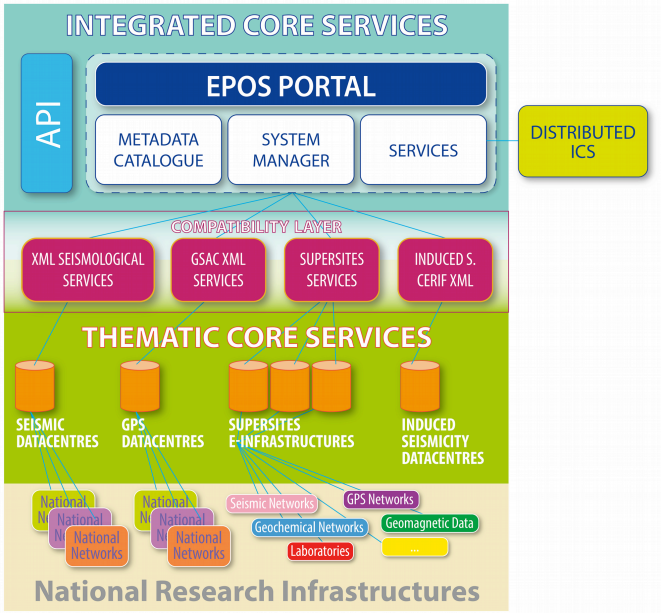

Figure 2: Functional Architecture, describing the technical functional components of EPOS. It specifies for each layer the ICT modules and their function. At the ICS layer it is fundamental to understand the design of the integrating e-Infrastructure.

Therefore, the ICS, being devoted to the real integration of data, data products and services, represent the "core" of the whole e-infrastructure, In their actual design, ICS are made up of several, modular, interoperable building blocks (top of Fig. 2):

1. *Metadata catalogue* is the key component. It contains all the information that the system might be willing to deal with. It uses CERIF[7] (Common European Research Information Format, recommended to the EU Member States) as a tool to harmonize databases on research projects The metadata describe datasets but also software, services, users and resources such as computers, datastores, laboratory equipment and instruments. It is clear the need for extensive maintenance of this information in the catalog. This can be done either by human or by automated means (recommended), depending on the technologies implemented at the TCS layer, to which the metadata catalogue will connect through the *compatibility layer*. This is a typical task accomplished by the System Manager.
2. *System Manager* can be considered as the "intelligence" of the system and is basically a software which manages the whole system. The System Manager takes advantage of the information contained in the catalogue (which is the "knowledge base" of the system) and makes proper decisions according to: (i) user requests, (ii) available resources, (iii) metadata contained in the EPOS metadata catalogue. Therefore, in a EPOS context, this is the place where the brokering techniques – but driven by the metadata in the catalog rather than by program code in the software – will be effective.
3. *EPOS Portal and API* are functional blocks dedicated to the interaction between users (both, human and machine) and the system. The former deals with the interaction between a human user and the system. A generic user will be enabled to perform actions like: (i) data/data products/sensors/facilities discovery, (ii) data/data products download, (iii) data/data products visualization, (iv) data/data products modeling and processing. However, this layer is not sufficient because EPOS wants to (i) be interoperable with other systems, (ii) be compliant with major European standards, (iii) deliver a high quality service that enable a user to perform programmatically some actions. A locus dedicated to machine-machine interaction is therefore

needed. This is exactly the Application Programming Interface (API), which includes a set of native functions enabling a machine use of the EPOS system, as for instance RESTful queries of the type: ***GET*** */entity?data_types=[seismwav,GPS,satdata]&lat=45.5345&lon=16.334&startime="datetime".*
For this latter purpose, the reliable and fully CERIF compliant CERIF-XML standard is used at the present stage in the form of a RESTful service that can be queried to obtain XML-formatted metadata.

4. *Services* interface module includes all the software and interfaces required to connect outsourced resources as, for instance, linkage to HPC centres or workflow management infrastructure (e.g. VERCE[16]).

4. Data Model

As already mentioned, the key to the architecture is the metadata catalogue and hence the metadata model. Metadata can be looked at in two dimensions:

- Metadata to describe the objects of the "EPOS ecosystem"
- Metadata for discovery and services

4.1. Metadata to describe EPOS ecosystem objects

This dimension of the metadata concerns the objects of the EPOS "ecosystem": these are classified into users, services (including software), data and resources (computing, data storage, instruments and scientific equipment) as shown in Fig. 3.

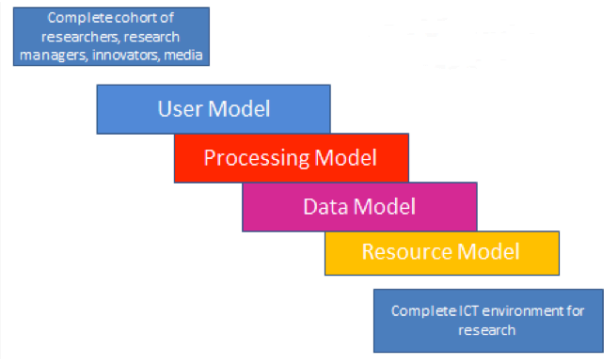

Figure 3: the four models to describe the EPOS ecosystem objects

The *User Model* describes how a subject (human user, but also a program, or a process) can interact with the EPOS e-infrastructure and determines the design of the EPOS web-portal. This is important to ensure all kind of people, regardless their location, language, expertise and disabilities, can easily access and use the system. Therefore, it will provide the technical instructions to ensure users' security, privacy and trust through its identification, authentication, authorization and accounting (IAAA). IAAA are based on the data policy and access rules describing the degree of openness of the information, data usability, data ownership, and the stakeholders' metric aimed at analysing the impact, influence, engagement, exchanges and ethical risks associated to each user category and the possible utilization of EPOS data and services

The *Processing Model* is by far the most difficult to create because it has to include all the instruction on how the system performs the calculations and visualization on the different data available within EPOS. It shall provide the core information of ICS, in particular the know-how on integrating data and data products beyond the simple data mining and data archiving currently scattered but still available at a community and national level.

The *Data Model* serves to describe the data and their associated metadata in order to allow the user to find them, work with them (integrate) and download them.

The *Resource Model* is a technical description of the physical resources owned by the data providers (i.e. national RIs and Thematic Core Services) that are available for EPOS integration and of those owned outside the EPOS delivery framework that will provide specific IT services (see the processing model). This model is needed to provide the description of the organization of the facilities in both their hardware and software components that will

guarantee: (i) data repository, (ii) data processing (calculations) and (iii) visualization (rendering). For each category the model pro-vides a detailed technical IT description with specifics on how to ensure a sustainable and efficient connectivity and therefore to allow the user to reach their content (the data) or to use them (processing and visualization).

4.2. Metadata for discovery and services

The other dimension of metadata aims at collecting all the information, which enables a user to perform actions and functions over data and data products. This model, the so-called three-layer model[8] is structured as follows (see also Fig. 4):

1. The discovery layer, using Dublin Core as metadata system extended to include the capability to generate from the underlying contextual layer – in addition to Dublin Core – DCAT, INSPIRE and both CKAN and eGMS to allow integration with government open data (data.gov) sources;
2. The contextual layer, using CERIF (Common European Research Information Format, recommended to the EU Member States as a tool to harmonize databases on research projects);
3. The detailed layer, which includes detailed metadata standards by domain or even individual database for each kind of data (or software, computer resources or detectors/instruments) to be (co)-processed.

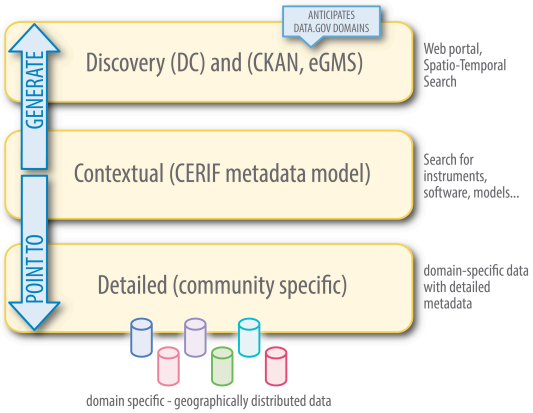

Figure 4: Three layer Metadata Architecture

4.3. CERIF model in the context of EPOS

The core of the proposed structure is the contextual layer, which is built following the concepts and guidelines of euroCRIS organization[19]. The proposed general scheme includes all the main actors and object which participate to the process of research, and which are represented as entities in the Entity Relational scheme. An extensive description of the model and the entities can be found in the Full Data Model documentation[7].

However such scheme was originally thought to describe all aspects of research process which ended up with the publication as a final product. Within the context of EPOS, the product of research is a more complex object, a data product which can potentially include very diverse kind of data and data products. To this extent some modifications were already done: in 2013 a Data extension proposal started by investigating CKAN, DCAT and eGMS and was guided by a draft proposal of the Jisc-funded CERIF for Dataset (C4D)[9] project.

As a comprehensive list of all the possible data products would have been very difficult, if not impossible, to draft, a categorization has been carried out taking into account previous work done from NASA, Interface Region Imaging Spectrography (IRIS), UNAVCO[10, 11, 12] and others. This is the so called EPOS data levels categorizaiton:

- **Level 0**: raw data, or basic data (example: seismograms, accelerograms, time series, etc)
- **Level 1**: data products coming from nearly automated procedures (earthquake locations, magnitudes, focal mechanism, shakemaps, etc)

- **Level 2**: data products resulting by scientists' investigations (crustal models, strain maps, earthquake source models, etc)
- **Level 3**: integrated data products coming from complex analyses or community shared products (hazards maps, catalogue of active faults, etc)

The modification proposed by C4D is able to handle all the data encompassed by this categorization.

4.4. Data discovery at EPOS ICS

The three-layer metadata structure can effectively manage the maximal level of commonality among all the metadata describing datasets provided by the data providers. The community specific metadata (lowest level) is hence not ingested by the system: only a subset of it is mapped into the Metadata Catalogue. However in order to have a reliable access to the local data, the Integrated Core Services had to set up efficient communication mechanisms into the so called *compatibility layer*. This layer makes possible the linkage between ICS and TCS thus enabling discovery and integration features. The thematic core services (TCS) are developed independently by their respective communities and in order to provide data and metadata to the ICS (but more in general to be interoperable at international level) they should provide software interfaces to access their systems, usually just end-user services to discover appropriate datasets or software and –in some cases – limited processing.

To fetch and discovery the desired data and metadata, ICS can then: (i) access to TCS web-services, (ii) access to TCS generic APIs, (iii) link directly to datasets and ingest the metadata either manually or by means of some automated process. The latter solution is highly un-recommended because of the limited flexibility and huge amount of work required, and is listed here just for the sake of completeness.

To enable such a communication (compatibility layer) a new entity was introduced in the CERIF scheme – the *cfServiceInterfaceDescription* – and the *cfService* was used with a special meaning: the entity is supposed to store information about the webservice or API providing data.

The purpose of these two entities is to store all the information which are necessary to enable the system to connect to the desired service and map the metadata of interest into the *cfResulProduct* entity.

5. Data Intensive Science

The aim of EPOS is to provide for the solid Earth community a research infrastructure for data intensive science[1] making use of integrated data. To achieve this, two clearly differentiated steps are needed: (i) integration, (ii) intensive data processing. The latter goes beyond the possibility of the EPOS Preparatory Phase Project and the e-infrastructure just depicted, whose role is to orchestrate the use of distributed processing facilities (and for instance determine whether it is convenient to move data to HPC centers or code to local repositories with some processing capabilities[20]). Therefore the former step is fully covered by the EPOS ICS. While some general purpose discovery tools, protocols and mechanisms, as for instance Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[21,22], can be effective for discovery of data repositories or of data itself, when one wants to select and collect huge amounts of data with a fine tuning over the parameters for selection and be able to mange their processing, then a robust handling of metadata is needed. Such tuning is often difficult when using common metadata standards (Dublin Core[23] is an example) alone, because they deal with generic metadata elements avoiding the heavy task of finding a semantic correlation across very different disciplines.

When heterogeneous data are confined within the field Earth Sciences, a general discovery can be carried on using the contextual layer of the just shown three-layers structure, which contains the maximal level of commonality among all the data products. Using the information contained into the metadata catalogue, the system can hence retrieve community specific metadata, thus enabling the user to perform a fine tuning of the discovery parameters (i.e. using specific metadata elements rather than generic ones) and the system itself to orchestrate higher level functions (visualization and processing) over distributed resources.

With such a robust management of metadata through a metadata CERIF-based catalogue, the path to data-intensive science gets closer and closer, thus creating new perspectives for science data processing.

6. Conclusion

In this work we outline the main concepts of EPOS, and describe its e-infrastructure devoted also to data intensive science. Such e-infrastructure makes use of the CERIF data model not for its usual domain of managing research information but to run a complex metadata engine which will enable EPOS services to perform advanced functions which will improve the capabilities of scientists dealing with Earth Sciences. EPOS has therefore demonstrated the power and utilization of the CERIF data model for building a data-intensive e-infrastructure for geoscience.

Acknowledgements

The authors acknowledge the work of their colleagues in EPOS WG7 e-infrastructure team, in particular the work of Luca Trani, Frieder Euteneuer, Damian Ulbricht and Alessandro Bartoloni, Carmela Freda from the EPOS Management Office.

References

1. Bell, G., Hey, T., & Szalay, A. (2009). *Beyond the data deluge*. Science, 323(5919), 1297-1298.
2. Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm*. *Data-Intensive Scientific Discovery*. Microsoft Research.
3. Halevy, A., Rajaraman, A., & Ordille, J. (2006). *Data integration: the teenage years*. Conference on Very Large Data Bases.
4. Basili, R., Valensise, G., Vannoli, P., Burrato, P., Fracassi, U., Mariano, S., Boschi, E. (2008). *The Database of Individual Seismogenic Sources (DISS), version 3: Summarizing 20 years of research on Italy's earthquake geology*. Tectonophysics, 453(1-4), 20–43.
5. Lenzerini, M., Sapienza, L., Salaria, V., & Roma, I.-. (n.d.). *Data Integration : A Theoretical Perspective*.
6. Suarez, G., Eck, T. Van, & Giardini, D. (2008). *The International Federation of Digital Seismograph Networks (FDSN): An Integrated System of Seismological Observatories*. Systems Journal, (3), 431–438.
7. Jeffery, K., Grootel, G. Van, Asserson, A., Dvorak, J., Rasmussen, H., Council, T. F. Republic, C. (2010). *CERIF 2008 - 1.2 Full Data Model (FDM)*.
8. Jeffery, K., Asserson, A., Houssos, N., & Jörg, B. (2013). *A 3-Layer Model for Metadata.* Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013, 3–5.
9. CERIF for Datasets home page: http://cerif4datasets.wordpress.com/
10. Nasa Data Processing Level: http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/
11. Interface Region Imaging Spectrography (IRIS) data levels definition url: https://www.lmsal.com/iris_science/doc?cmd=dcur&proj_num=IS0076&file_type=pdf
12. UNAVCO data levels definition url: http://pbo.unavco.org/data/gps
13. EUDAT website for further information: http://www.eudat.eu/
14. GEOSS website for further information https://www.earthobservations.org/geoss.shtml
15. EPOS website for further information http://www.epos-eu.org
16. VERCE website for further information http://www.verce.eu/
17. ORFEUS website for further information http://www.orfeus-eu.org/
18. Doan, A.H. and Halevy, A. and Ives, Z. , *Principles of Data Integration*, Elsevier Science 2012
19. ORFEUS website for further information http://www.eurocris.org/
20. Aloisioa, G., Fiorea, S., Foster, I., & Williams, D. (2013). *Scientific big data analytics challenges at large scale*. Proceedings of Big Data and Extreme-Scale Computing (BDEC), 2–4.
21. Devarakonda, R., Palanisamy, G., Green, J. M., & Wilson, B. E. (2010). *Data sharing and retrieval using OAI-PMH*. Earth Science Informatics, 4(1), 1–5. doi:10.1007/s12145-010-0073-0
22. Khan, Nadim Akhtar. "Emerging Trends in OAI-PMH Application." *Design, Development, and Management of Resources for Digital Library Services.* IGI Global, 2013. 147-159. Web. 15 Apr. 2014. doi:10.4018/978-1-4666-2500-6.ch013
23. Weibel, S. (1997), *The Dublin Core: A Simple Content Description Model for Electronic Resources*. Bul. Am. Soc. Info. Sci. Tech., 24: 9–11. doi: 10.1002/bult.70