Report from 9th RDA plenary in Barcelona 3rd-7th April

By Bo Weidema & Michele De Rosa, BONSAI

Topics:

What is RDA?

The Research Data Alliance (RDA) is a community-driven organization founded in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. RDA members currently meet in plenaries held twice a year.

RDA aims at building the social and technical infrastructure to enable open sharing of data. With more than to 5,400 members from 123 countries, RDA provides a neutral space where its members can come together through focused global Working Groups (WGs) and Interest Groups (IGs) to develop and adopt infrastructure that promotes data-sharing and data-driven research, and accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographical and generational boundaries.

Members can propose Birds of a Feather (BoF) sessions at the plenary meetings to discuss topics not yet endorsed by RDA, which may or may not result in the formation of a WG or an IG.

Mark Parsons, Secretary General of RDA, stated that "*We* [RDA] *do not solve challenges. We enable the challenges to be addressed*".

For a general introduction to the state of the art of the field of virtual research infrastructures, see also the Annex to this report.

Current RDA recommendations and their relevance for BONSAI

- "Data Citation of Evolving Data" is an output of the Data Citation Working Group. Bo met with co-chair Andreas Rauber. The recommendation is relevant for BONSAI (see below on "Versioning and Citation").
- The Repository Audit and Certification DSA–WDS Partnership Working Group has produced a catalogue of common requirements for certification of data repositories (databases). This can be relevant initially as a checklist for self-declaration of BONSAI, and eventually for certification. Many of the requirements will only become relevant once BONSAI holds data in a form that is unique (e.g. due to its origin or curation) and therefore has a value that will motivate data suppliers to ask BONSAI for its certification. The requirements will be less relevant if data will be stored in a distributed form or in cooperation with another entity that is already certified.

Data service providers

There seems to be a large amount of service providers for data management:
We have already considered CKAN. Below is an annotated list of the ones we came across at RDA, in an attempt at identifying discriminating characteristics:

- EOSC (European Open Science Cloud)
- EUDAT (European Data Infrastructure) and their B2 Service Suite
- INDIGO-DataCloud
- EGI
- CKAN
- D4Science
- DuraSpace
- ELEXIR

From a BONSAI perspective, it is not very clear what distinguishes EUDAT, INDEGO, EGI and CKAN and they all seem to cooperate and offer somewhat parallel entry points. We should probably take contact to all of them to understand better how we can take advantage of their offerings. In July Michele will be participating to the EUDAT Summer School to understand better EUDAT (see below) functions and potential applications.

EOSC (European Open Science Cloud)

The EOSC is a European high-level initiative aiming to accelerate and support the current transition to more effective Open Science and Open Innovation in the Digital Single Market. EOSC is intended as a federated environment for scientific data sharing and re-use, based on existing and emerging elements in the Member States, with light-weight international guidance and governance, and a large degree of freedom regarding practical implementation. Read more at https://eudat.eu/european-open-science-cloud. The EOSC is currently in its very first initiation stage, with a pilot project starting from January 2017 to be finalised by 2020.

EUDAT (European Data Infrastructure) and their B2 Service Suite

The EUDAT project is funded by the EU Horizon2020 program, started in 2011, now in 2nd phase EUDAT2020, and will end 2018. Beyond 2018 there is an agreement of an EUDAT-EGI-Indigo consortium on the EOSC-Hub proposal. Over 20 major European research organisations, data and computing centres have signed an agreement to sustain the EUDAT for the next 10 years giving the birth to the EUDAT Collaborative Data Infrastructure (or EUDAT CDI). EUDAT partner organisations are fundamental to the realisation of the EOSC (see above) and many of them are involved in the EOSC pilot project, tasked with the responsibility to design the EOSC.

The EUDAT Collaborative Data Infrastructure (CDI) is a defined data model and a set of technical standards and policies adopted by European research data centres and community data repositories to create a single European e-infrastructure of interoperable data services. The EUDAT CDI is realised through on-going collaboration between service providers and research communities working as part of a common framework for developing and operating an interoperable layer of common data services. The scope of the CDI covers data management functions and policies for upload and retrieval, identification and description, movement, replication and data integrity.

Using the EUDAT CDI, researchers can “rely on innovative data services to support their research collaboration and data management”. In particular:

1. Benefit from a simple, efficient, trustworthy, affordable collaborative and interoperable data infrastructure
2. Connect with Europe’s most powerful supercomputers
3. Be supported by a pool of experts to tackle data challenges
4. Trace data origin in the metadata catalogue

The services available for users can be divided into three main areas as follows:

- Data Access and Re-use, covering data storage with B2DROP (dropbox-like personal cloud for storing long-tail research data during on-going research collaborations, not persistent) and B2SHARE (primarily for storing very large sets of data) with the annotation service B2NOTE, data search with B2FIND (searching based on metadata descriptions of data stored in EUDAT data centres and in other data repositories) and authentication and authorization of users with B2ACCESS (enables users to log in using various identities – for example, an identity from the research organisation they work for or with a Google account).
- Data Preservation, covering support for implementing data management policies (e.g. for data replication, and data integrity checks) with B2SAFE (the B2SAFE Data Policy Manager, which is currently under development, will enable data managers to manage policies centrally or locally), and providing persistent identifiers (PIDs) with B2HANDLE
- Data Processing with B2STAGE (transferring sets of data that are usually very large)

A complete EUDAT Service Catalogue is available at this link, except for B2NOTE, which is currently at pilot stage. The services are generally free for European researchers.

INDIGO-DataCloud

INDIGO is a project funded under Horizon2020. INDIGO-DataCloud delivers open source software components tailored to scientific communities and to e-infrastructures, aimed to increase ease of use and effectiveness in the exploitation of High performance computing (HPC) clouds, and thereby help overcome current challenges in the cloud computing, storage and network areas.

The project will extend existing PaaS solutions, allowing public and private e-infrastructures, including those provided by EUDAT (see above), to integrate their existing services and make them available through Authentication and Authorisation Infrastructure services compliant with GEANT’s interfederation policies, thus guaranteeing transparency and trust in the provisioning of such services. INDIGO will also provide a flexible and modular presentation layer, allowing innovative user experiences and dynamic workflows, also from mobile appliances. INDIGO framework or services intends to have a low learning curve, based on open source software, rooted in a large number of use cases, exploiting available, general solutions rather than custom-made specific tools or services. A short intro is available at https://www.indigo-datacloud.eu/node/425.

The INDIGO Service Catalogue provides links for access and download of a large number of software components, including the *data management system* ONEdata, all documented under this heading.

EGI

EGI is a foundation based in Amsterdam that delivers advanced computing services to support scientists, multinational projects and research infrastructures. Their service catalogue and user cases appears to focus mainly on services for big data, but currently a new EGI Applications on Demand service platform is run in a beta version that allows individual researchers and small research teams to perform compute and data-intensive simulations on large, distributed networks of computers. This service is under development while available for testing publicly. The EGI Federated Cloud is a cross-domain federation of existing academic private clouds and resource providers, built around open standards. The EGI Open Data Platform is a solution allowing integration of various data repositories available in a distributed infrastructure deployed into the EGI Federated Cloud using ONEdata (mentioned above under INDIGO) as core enabling technology. There is no upfront cost.

CKAN

The CKAN association is an autonomous association hosted by the Open Knowledge Foundation. CKAN is an open source data portal and data management system. The core functionality of CKAN provides “a wealth of features and has over 200 community extensions which can fill almost any feature gap”. CKAN is used by EUDAT for their B2FIND function.

D4Science

D4Science.org is a customised CKAN infrastructure, enabling Virtual Research Environments, each with their own metadata format, powered by gCube software to publish their data in DCAT, RDF and JSON formats, searchable through the D4Science integrated data catalogue. D4Science’s service catalogue offers a rich array of services to its end users directly or to Infrastructure Managers and Service Providers under three exploitation models.

DuraSpace

DuraSpace is a non-profit corporation that provides “leadership and innovation” for open technologies that promote durable, persistent access to “the world’s digital heritage." They have three open source projects: DSpace - a document repository application, Fedora - a data repository application, and VIVO - a software and an ontology for representing scholarship. DuraSpace also provides three services: DuraCloud, DSpaceDirect and ArchivesDirect - all services focussed on low-cost, cloud-based archiving, discovery, access and preservation. For BONSAI, the most relevant application is probably Fedora.

ELIXIR

ELIXIR is similar to EUDAT but specialised for the life sciences research community. It collaborates with e-Infrastructures, such as GEANT, EUDAT or EGI. ELIXIR have a number of services run by the national centres. For example, the Dutch ELIXIR are developing Fairifier, Fair data point, Fair data, Orka (open data knowledge annotator) which are parallel to EUDATs: - ; B2share; B2find, B2note.

Big Data Stakeholders Meeting (Interest Group)

Michele attended the Interest Group session on Big Data. Peter Baumann (one of the world leading scientists in data science) is chair of the Interest Group. Samuel Kerren (EPFL Lausanne - in contact) presented "*A semantic provenance-based data platform to support large-scale data discovery and integration*" tailored on neuroscience (Blue Brain Project) but an interesting example that could be transferred to our scientific domain, e.g. inspiring a similar architecture for BONSAI: the amount of data neuroscience draws from and the structure of the research model they follow (data gathering, model building, simulation - supercomputing too in their case - analysis, validation, model refinement) is very similar to the LCA framework. The problems they face are also similar:

- Integrate data (ranging from small to PB DBs)
- Need to discover similar and related data
- Need for data provenance (tracking origin, reproducibility, tracking dependencies), using W3C-PROV)

Their objective is to make their platform open source (2-3 months from April 2017) and production ready, provide API to interface all aspects of the platform and UIs (to edit, browse and discover data) and to build a community of users.

Currently the Blue Brain Nexus approach is very generic but they intend to 'bend' it towards specific scientists' domain. Thus, there may be a potential for future collaboration and shared interest.

Legal issues

During the social dinner, Bo met a lawyer from Bern, Dr. Willi Egloff, that has a firm (advocomplex.ch) dealing with copyright issues for data harvested from scientific literature. He also has an NGO that deals with the actual harvesting.

FAIR + R + principles

Bo attended part of a workshop on FAIR arranged by the Bluebridge project consortium, which supports capacity building in interdisciplinary research communities within fisheries and aquaculture, using the services of D4Science.org (see above).

The FAIR principles:

- Findable
- Accessible
- Interoperable
- Reusable

were originally published in Nature and currently being maintained as a living document.

The principles are supported by main data service providers, such as EUDAT.

Some add an extra R (for Reproducibility) and more other issues such as security/legal issues may also be added. It is relevant to see this in a sound data life cycle perspective, from raw data over processed data, curated data, to analysed data, covering also time of publication and standards for preservation.

However, FAIR does not imply Open, nor Semantic web integrated. The concepts focus primarily around transparency, exposing data in multiple formats, ensuring traceability of origin, citation, etc.

Metadata harmonisation and classification issues

A lot of the current efforts appear to be focused on discoverability and interoperability, for which harmonisation of metadata and classifications are central issues.

Bo attended two meetings of the RDA WG on data type registries. Data type registries seek to facilitate parsing, automated selection, interpretation, and processing of large amounts of scientific data without human intervention, which is particularly relevant especially across domains when domain specific knowledge is not available. The term 'data type' is the characterization of data structure at multiple levels of granularity, from individual data points up to and including large data sets. Optimizing the interactions among all of the producers and consumers of digital data requires that those types be defined and permanently associated with the data they describe. Further, the utility of those types requires that they be standardized, unique, and discoverable. Simply listing and describing types in human readable form, say in one or more open access wikis, is certainly better than nothing, but full realization of the potential of types in automated data processing requires a common form of machine readable description of types, i.e., a data model and common expression of that data model. A prototype is available at typeregistry.org

An example of a data type for a measurement output for a stream gauge could look like this:

ID
name: stream gauge:
Description
Standards: ISO name: no. …
Properties…
etc.…

Use case:
JSON schema for data type model
From spreadsheet or rdf to Cordra repository (Cordra is an open source Digital Object Repository and Registry software).
It is an open question where to deploy the registry.

Use case:
Enrich.cordra.org select tabular datasets from data.gov are harvested and processed in an automated fashion for demonstrating the value of adding descriptive details useful for consumers using the cordra software (mentioned above). That automated process resulted in
1) identifying the column names in those datasets, and
2) identifying the syntactic nature of the values in each of those columns, e.g., that column 1 is an integer and the values range between 100 and 5000, that column 2 is a date and the date format is yyyy-mm-dd, and so on.
All the generated descriptions are "data type records". Those data type records can be manually edited or enhanced even further. For example, semantic information such as that column 1 is not just an integer, but is also a "temperature expressed in fahrenheit" could also be stated. The Enrich service use a "data type registry" behind the scene to store "data type" records, having registered several hundred concepts (e.g., pressure, temperature, length, etc.) and measurement units (e.g., Fahrenheit, Celsius, etc.). The metadata records that data.gov provided along with the datasets have also been harvested and is stored in a metadata registry. In order to link datasets, data type records, and metadata records together, handles (from the Handle System) are used. Currently, the data type records are represented using JSON and can easily be represented using RDF. The data type record structure does not conform to any existing standard, but relevant efforts are ongoing to evaluate existing standards. A pdf slideshow is available that explains the achievements nicely.

Use case from climate data processing:
Concept:
netCDF-files
Collection
<metadata> xml
third-party input
to
Processing service

Use case: AgGateway

- no controlled vocabularies
- 240 companies dedicated to data exchange standards
- semantic assets: representations of universal variables; geopolitical context; observation codes
- geopolitical context problem: epa-number; bundesorten amt, tax data,…)
- contextItem:
- ISO 19135

Use case: ePIC
dtr.pidconsortium.eu
150 basic info types
89 info types
168 types
type life cycle: in preparation, candidate, approved, deprecated
Migration to candidate is non-trivial due to dependencies, and the hierarchical concept that all subtypes then need to be candidates as well
SI units and derived units (e.g., Joule)
constants as types? (kilo = *1000); Types are already variables...

ISO-IEC/JTC1/SC32/WG2 is working on a standard ISO 11179-7 for a metamodel for dataset description. However, this is about datasets rather than internal details of a dataset (data type records). Other relevant standards are ISO 11179-3; ISO 19763-12 and ISO 11404. The Corporation for National Research Initiatives (CNRI) will create a UML diagram referencing pieces from these standards. This should be seen as a strawman ("strawman" is jargon in american business and software development language used to represent a proposal that is not meant to be seen as final, but intended to provoke improvements and thus be destroyed).

Now the target for the WG is to connect data types for datasets across infrastructures based on co-authorship or other collaboration such as joint funding or grants. This will be represented in a neo4j graph database. The Global Research Identifier Database was mentioned in this context.

The Australian & New Zealand classification of socio-economic objectives was presented. This classification allows research to be categorised according to the purpose or outcome of the research as perceived by the data provider (researcher). It consists of discrete economic, social, technological or scientific domains for identifying the principal purposes of the research. The classification is available through Research Vocabularies Australia which is a repository for vocabularies for research.

Bo attended a workshop on PID kernel information. The idea of this Working Group is that certain metadata should be linked to the PID and reside within the PID resolution system itself, in order to speed up Internet search. However, since this "kernel information" is not domain specific, it is hard to see how this can be used to distinguish potentially useful data from the not so useful, which should be the purpose of the query.

For classifications in general, it should be noted that when classifications evolve, a class can be disaggregated, so that its new representation will be as an aggregate of (links to) two or more sub-classes each with their name and PID.

Michele attended the Metadata joint session. The objective is to create a new catalogue to improve the quality of metadata (thus of the data) and to facilitate the search and the automatic access. They have a github page for the group and a page tracking the development of the catalogue. However, our scientific domain (industrial ecology) has so far been completely out of the discussion and there are no case study uploaded showing examples of our metadata current structure (e.g. ecospold2). Michele is in touch with Rebecca Koskela (the DataONE Executive Director and one of the group chairs) to upload one. Michele and Rebecca are also discussing the possibility to propose a BoF inviting and coordinating the participation of the industrial ecology (see also below under “Follow up”).

Michele also attended the joint session on material data and infrastructure. It was interesting but, as a fusion of old interest groups, not really easy to follow. Michele talked to Ray Plante from NIST that shared the NIST’s GitHub page for the development of resource metadata schemas and related tools in support for the Materials Genome Initiative at NIST.

Ontologies and semantic web issues

Bo attended several meetings on semantics and ontologies, including the RDA Interest Group on Data Foundations and Terminology (RDA-DFT), the EUDAT Semantic Working Group, and the Domain Vocabulary BoF, the latter being a new group intended to assist standardising the registration and management of domain vocabularies. Bo will follow the work of this BoF. Next physical meeting in Montreal. For BONSAI, Pascal & Michele may attend.

The discussion centres on principled ontologies, referring to the principles of the OBO foundry. A general wish list for ontologies is that they should allow interoperability, validation, and flexibility for adding new concepts.

The Extensible Observation Ontology (OBOE) is a formal ontology for capturing the semantics of scientific observation and measurement. The main concepts in OBOE include:

- Observation: an event in which one or more measurements are taken
- Measurement: the measured value of a property for a specific object or phenomenon (e.g., 3.2)
- Entity: an object or phenomenon on which measurements are made (e.g., Quercus rubrum)
- Characteristic: the property being measured (e.g., VolumetricDensity)
- Standard: units and controlled vocabularies for interpreting measured values (e.g., g/cm^3)
- Protocol: the procedures followed to obtain measurements (e.g., DensityProtocol2014)

OBOE can characterize the context of an observation (e.g., space and time), as well as dependencies such as nested experimental observations. It includes an extensive set of unit definitions (e.g., grams of carbon per liter of seawater), and can facilitate automatic unit conversions (e.g., pounds to kilograms). OBOE can be easily extended to create Entities and Characteristics for specific research domains, making it both broadly applicable and highly customizable.

As a basic introduction to ontologies, “A guide to develop your first ontology”, was mentioend as “necessary and sufficient” obligatory reading (!).

Linked open data is seen as the most important driver for the development of ontologies. Linked Open Vocabularies has an upload interface that to some extent curates ontological metadata. Annotating and archiving ontologies were mentioned as use cases.

The RDA-DFT has a Term Definition Tool, called TeD-T, to support broader model and vocabulary agreements for and across the RDA WGs and IGs along with their representative communities and stakeholders. One can add own terms and populate fields such as name, definition, explanation, example, sources, etc.

Several other related initiatives were mentioned:

- The CASRAI dictionary is a pilot to demonstrate the value of a dedicated set of terms for the research data domain.
- SKOS under W3C develops specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web).
- Cordra software for schemas.
- Reptor PHP application which demonstrates functionality of a modern data repository offering untyped metadata handling by supporting several types of serialised metadata (plain text, key-value pairs, XML formats), PID handling independent of the used PID system (DOI, Handle, URN etc.), access to the object data (bitstreams) stored on the filesystem and taking care of permission management, Integration of Data Type Registries, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and ResourceSync (partially), and social media integration.
- Swoogle search engine for Semantic Web ontologies, documents, terms and data.
- Bioportal (managed by John Graybeal from Monterey Aquarium) which from a free text query can recommend suitable ontologies.
- Biosharing.org
- Schema.org
- Pundit semantic annotation
- The Center for Expanded Data Annotation and Retrieval (CEDAR) develops information technologies that make authoring complete metadata much more manageable, and that facilitate using the metadata in further research. Based in the biomedical field.
- The Centre for Ecology & Hydrology (CEH) provides access to different vocabularies within geological, ecological and hydrological sciences, including envthes.

- The PolisGnosis project, which develops ontologies for representing the complete definition of each indicator in the ISO 37120 standard on “Sustainable development of communities -- Indicators for city services and quality of life”, representing each city’s indicator value (for a particular year), including the supporting data used to derive it, using the aforementioned ontologies, further develop ontologies for representing indicator theme-specific knowledge, such as basic knowledge about education, such as school, teacher, student, grade, etc. and represent a city's theme-specific indicator knowledge, develop axioms to determine the consistency of indicators, and develop a reusable, interoperable logical theory of longitudinal and transversal analysis of metrics. There is a slide presentation on the project. Two new standards ISO/IEC AWI 21972 “Information technology - An upper level ontology for smart city indicators” and ISO/IEC AWI 30145 “Information technology - Smart city ICT reference framework” is under development. This ontology is probably closer to the technosphere modelling of BONSAI than any of the other ones presented here.
- Ontology Summits, an annual series of events that involves the ontology community and communities related to each year's theme chosen for the summit.

Versioning and citation

Nice presentation by Andreas Rauber on the output from RDA WG on “Data Citation of Evolving Data”. The recommendation document contains 13 specific recommendations, with the aim to allow identification, citing and retrieval of arbitrary views of data, from a single record to an entire dataset, as it existed at a certain point in time, in a precise, machine-actionable manner, that is stable across different technologies and technological changes. The basic idea of the recommendation document is that all that is needed is to ensure timestamping of each change to individual data points and each query (and query output). An extensive description of the recommendations has been published in the bulletin of the IEEE-TCDL (Technical Committee on Digital Libraries) and there are also a number of webinars explaining both the recommendations in a bit more detail, as well as adopter reports, all linked to from the webconference site.

For BONSAI, the recommendations (R1-13) imply the following specifications:

- Ensure that any additions, edits and deletions on individual data points and datasets are marked with a timestamp, which at the same time works as version information (R1&2)
- Ensure that data points in dataset can be sorted unambiguously and reproducibly (R5)
- Normalise queries and compute a checksum of the normalised queries so that identical queries can be detected efficiently (R4)
- Assign a new PID to new queries and store queries and the associated metadata in order to re-execute them in the future (R3, R8&9)
- Provide checksum of query results to enable verification of the correctness of a result upon re-execution (R6)
- Mark query results with a timestamp based on the query execution time or the last update to the database prior to the query execution time. This allows retrieving the data as it existed at the time a user issued a query (R7)

- Assign a new PID to each new query result, and assign the existing identical PID to identical query results, i.e. query results with identical checksums, with the same query PID, performed on an unchanged database (R8)
- Lower the barrier for citing the data by generating citation text snippets for query results, including the query PID and query result PID (R10)
- Make the PIDs resolve to a human readable and machine actionable landing page that provides the data (via query re-execution) and metadata, including citation text snippet and the PID of the database (R11&12)
- Add to data management policy that when data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), also the queries and associated checksums are migrated, and the migration verified, ensuring that queries can be re-executed correctly (R13&14).

Michele attended the session on 'How to give credit to scientists for their involvement in making data & samples available for sharing'. Rodrigo Costas (CWTS Leiden) showed the results of a study "The value of Research Data" (Knowledge Exchange 2013) and the practical possibilities based on DataCite, a non-profit promoting open research data accessibility and tracking, advocating the use of DOI. DataCite is currently the largest data source on registered data publications. Michele also has the contact of a guy working at DataCite. One of the major obstacle so far (also shown by the research cited above) is the lack of completeness (e.g. common metadata for all records) and relevant missing elements, such as the affiliations of data creators (institutional and geographic data), organizing the concept of data (what type of data), info on dates of production and info on modes of data production (all often lacking in Datacite).

Text and Data Mining (TDM) policies and solutions

Michele attended the BoF on TDM. Generally speaking, TDM is of great interest in the long-term prospective for BONSAI. Yet, we are still far from having to tackle this issue. It is worth, however, to keep track of the development and be in contact with the people that are working on it:

- Stelios Piperidis, Head of Language Technology applications Department at the Institute for Language and Speech Processing, Athens. He presented the OPENMINTED project, an ongoing project testing technical solutions to TDM problems. (Michele is in contact with Stelios)
- Marco Caspers, researcher at the University of Amsterdam Institute for Information Law, where they are active on setting the policy framework (most of the obstacles are currently of political/"legal" nature) for mainstreaming the use of TDM tools. Marco spoke about the Future TDM project that seeks to improve uptake of TDM in the EU by actively engaging with stakeholders such as researchers, developers, publishers and SMEs. The objective of the project is to overcome existing legal, technological and skill barriers that prevent TDM technology from being adopted within the EU. Marco looked at our bonsai.uno page and asked Michele to write a guest blog on their platform on the impact of BONSAI on (future) TDM possibilities. The blog post is now available here.

For text mining with Phyton, NLTK is recommended, see also here.

Wikimedia integration

Bo met Mr. Luca Martinelli, secretary of Wikimedia Italy, and asked him about integration of data into Wikiboxes. He suggested to read on Wikidata's pages on Data import Hub and Data donations. The first issue would be to provide an identification of a wikipage as being about a product (or an activity) (see also how this is done by productontology.org *without* a Wikimedia integration), so that we can link these page names to our database names. Next issue would then be the creation of a box.

Software Source Code focus group

Michele attended this BoF and is in contact with the main speaker and proposer of the BoF (Roberto Di Cosmo senior scientist at the INRIA Paris, France biggest IT centre, and director of www.softwareheritage.org). Minutes of the whole meeting can be found here and the slides here. The Bof is very likely to be endorsed by RDA secretariat and was very successful. The focus was on the importance of source code and on taking care of it in the sense of storing it (including history) safely and making it readable and useful (well explained by this great quote by Harold Abelson: "*Programs must be written for people to read, and only incidentally for machines to execute*"). The BoF was inspired by the fact that progresses on Open Access Data Repositories and Open Data Set Repositories are occurring but there is currently no Open Source Repository for storing and finding code (Github as other similar websites can be shut down anytime as happened to GoogleCode web page and do not provide anyway consensus minimum requirements for the code to be interoperable and findable; also, the content can be modified and the history deleted).

Crediting scientists for sharing data

Michele attended the BoF meeting. The group discussed solutions to break the vicious circle trapping data sharing: no sharing is done because no reward is obtained; data sharing and reuse is therefore low, making data metrics difficult; lack of metric means that data sharing and reuse cannot be measured making it difficult to reward. A reward system should include data metrics and standardisation, formalisation and tracking of data publications and citations. It is proposed to use DataCite to analyse data production https://search.datacite.org/
DataCite is an international non-for-profit consortium (since 2009) of public research institutes, funding bodies and publishers worldwide. The mission: to promote open research data accessibility and tracking, advocating for the use of Digital Object Identifiers (DOI). DataCite is the largest data source on registered data publications. In April 2016 more than 7 million records were in DataCite. Most of them were data sets, followed by text and images. Using bibliometrics from Datacite source is a practical solution but some improvement is needed.

The outcome was the idea to write a paper starting with the input of this session and get input from other groups or from other community, or to give input in others. The attribution metadata working group share similar issues. The group meets virtually on May 31st.

Use cases from related science fields

SEDAC, the Socioeconomic Data and Applications Center, is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration. Focusing on *human interactions in the environment*, SEDAC has as its mission to develop and operate applications that support the integration of socioeconomic and earth science data and to serve as an "Information Gateway" between earth sciences and social sciences. This was one of the more interesting posters, and looks like something we should look more into for both data, tools and cooperation.

Gesis a discovery service in beta version based on up-to-date metadata that are harvested from social science research data collections worldwide. Together with the above, this was one of the few *social science and economic research* related presentations.

IndexMeed Consortium (Indexing for Mining Ecological and Environmental Data) using graph-based models to consider biodiversity data at different quality and heterogeneity, despite their differences, at a similar level, and improve decision support using emerging data mining methods (collaborative clustering, machine learning, mining graphs, knowledge representation, etc.).

Blue Brain Project http://bluebrain.epfl.ch/ and https://en.wikipedia.org/wiki/Blue_Brain_Project : The Blue Brain Project is an attempt to reverse engineer the human brain and recreate it at the cellular level inside a computer simulation. Goals of the project are to use biologically-detailed digital reconstructions and simulations of the mammalian brain in order to gain a complete understanding of the brain. This would allow to identify the fundamental principle of brain structure and functions in health and disease, thus enabling better and faster development of brain disease treatments. Data about all the many different neuron types are collected used to build biologically realistic models of neurons and networks of neurons in the cerebral cortex.

Next RDA plenaries

Next RDA plenary will be held in Montreal, 19-21st September (with co-arranged meetings around). See www.rd-alliance.org/rda10. We should consider if Pascal should attend, and whether Michele should also attend, the latter particularly if this could be combined with a working stay at CIRAIG focussing on concrete topics and deliverables in cooperation with the CIRAIG team.

11th plenary will be held in Berlin, March 2018.

Follow up

To be done: Integrate relevant points from the above (and from the annex) into the specifications on the BONSAI wiki.

Out of the 620 participants at the 9th RDA Plenary, the large majority were natural scientists and data scientists. One a small group of social science people participated, and these were mainly from sociology and archaeology. Apparently none, or very few, from economics and none from industrial ecology, although the data sharing issue for these sciences have many parallels to those of the other sciences. BONSAI may have an important role to play as ice-breaker for these communities to be involved in RDA and the open data world in general.

We should propose a BoF in the future. We already did it for this plenary but probably we were not sufficiently aware of the purpose and format. A BoF theme could be the creation of a new industrial ecology and economics science domain IG.

Michele attended the BoF on Disciplinary Interoperability Framework (DIF). The BoF is likely to become an IG with the objective of creating new opportunities for interaction between practitioners from across scientific and technological domains. Since our scientific domain is not yet well represented and we do not have a thematic IG, we should follow the development of this BoF. Most of what will be discussed here is probably of interest to our domain too.

Bo intends to follow the Domain Vocabulary BoF.

Annex: Virtual research infrastructure (RI) technology review

The most recent review of the current RI technology scene was made in 2016 by ENVRI. The below is a short summary of relevant findings for BONSAI on the topics of

- Data identification, authenticity, and citation
- Data curation and preservation
- Cataloguing
- System and workflow architecture, management and deployment
- Provenance
- Optimisation
- Semantic linking

Whenever possible, we should apply and/or link to such general recommendations and standardised nomenclatures and ontologies.

Data identification, authenticity, and citation

Good practice is to assign a persistent identifier (PID) to each digital object (data) to make data searchable, attributable, and citable.
Metadata is also a digital object.

Collections of data (datasets, databases) are digital objects that explicitly identify the data they are composed of, e.g. by specifying the ranges of a data query.
It is implicit that a PID needs to be unique and resolve to a locator for the resource (a URI). The Internet Engineering Task Force (IETF) has published good practice for URI Schemes.

Ensuring authenticity: Unintentional errors in data transmission can be detected by the addition of a Cyclic Redundancy Check (CRC), while detection of intentional adulteration requires public key cryptography.

For citation principles, reference to the FORCE 11 data citation principles is relevant, although these principles do not specify any actionable technologies.

Data curation and preservation

Data curation is the organization and integration of data collected from various sources, the addition of metadata (such as annotation, quality assessment, review, mapping to other data, and facilitating publication), so that the value of the data is maintained over time, and the data remains available for reuse and preservation. Data curation relates to data management plans, including policies on what should be curated (e.g. also how data are used and by whom?), the responsibilities in the curation process, and how the curation system itself may be accessed, used, and altered.

It is good practice to ensure that calculated data remains re-calculable with identical result, which requires that also software algorithms needs to be sufficiently documented and preserved, including the environments they need to be operational.

A good set of principles for preservation are provided by LOCKSS.

It is good practice to ensure that PIDs are resolvable, also when curation has ended, implying the placement of an informative tombstone for data that are removed.

Cataloguing

Cataloguing may refer to the use of external catalogues that are useful for a broad range of applications, and internal catalogues that are maintained within a database or domain.

The currently most relevant example of an application area for an external catalogue is for persons and organisations, for use in the context of data provenance. The most popular system for person's identification and cataloguing is currently ORCID, which is also working on organisation cataloguing. Research organisations are also catalogued by GRID, distinguished by a unique identifier, GRID ID. ORCID is involved in the THOR project, which defines relations between contributors, research artefacts (including data), and organizations to incorporate these relationships into the ORCID and DataCite systems, embedding new PID resolution techniques into existing services to support seamless direct access to artefacts, and in particular data, creating services to allow associations between datasets, articles, contributors and organisations at the time of submission, delivering the means to integrate trans-disciplinary PID services in community-specific platforms, focussing on cross-linking, claiming mechanisms and data citation.

Internal dataset catalogues for many research infrastructures are managed with CKAN (used by EUDAT for the B2FIND function) and Geonetwork (an open-source catalogue application software for spatially referenced resources and especially datasets). Both applications are web servers and can be used and managed online. It appears to be pragmatic and feasible to harvest existing CKAN and geonetwork catalogues in one CKAN central server.

For software, it should be possible to apply the gitHub API to harvest a catalogue of information related to software and algorithms.

Internal catalogues are relevant for all metadata classification issues, see the separate heading "Metadata harmonisation and classification issues".

System and workflow architecture, management and deployment

Scientific workflow management systems provide graphical user interfaces for creating workflows to facilitate the description of workflows during workflow development. Their output also serves as a basis for documenting workflows and facilitating sharing and reuse of workflows and code. Some notable systems are Dispel4Py, Triana, Apache Taverna (strong focus on web service invocation and integration of distributed and possibly heterogeneous data), KNIME (strong focus on user interaction and visualisation of results with focus on workflows from the fields of data mining, machine learning, and chemistry), Kepler (with major focus on statistical analysis), Pegasus, and WINGS. From the sheer amount of different systems it appears that there is currently no harmonised or integrated framework for these. Most of the systems have their own graphical workflow modelling and language. Obviously, the lack of standardized syntax and semantic description for workflow modelling and language results in many replicated works. More effort is thus needed towards workflow modelling standardization.

The Unified Modeling Language (UML) is the current standard for analysis, design, and implementation of software-based systems as well as for modeling business and similar processes. There is also a standard workflow language for web services BPEL, but without standard graphical notation. It is possible to translate UML graphs into BPEL documents for concrete workflows. An interesting attempt at combining different standards into a workflow ecosystem with strong links to provenance (see separate heading below) is provided by WEST (Workflow Ecosystems through STandards).

Most complex system specifications are so extensive that no single individual can fully comprehend all aspects of the specifications. Furthermore, different individuals will have different interests in a given system and different reasons for examining the system's specifications. A business executive will seek different information from a specification than an information scientist. The Reference Model of Open Distributed Processing (RM-ODP) therefore introduces the concept of viewpoints. Each viewpoint is a subdivision of the specification of a complete system, established to bring together those particular pieces of information relevant to some particular area of concern. Associated with each viewpoint is a viewpoint language that optimizes the vocabulary and presentation for the audience of that viewpoint. Although separately specified, the viewpoints are not independent. Key items in each are identified as related to items in the other viewpoints. Moreover, each viewpoint substantially uses the same foundational concepts as defined in RM-ODP. However, the viewpoints are sufficiently independent to simplify reasoning about the complete specification. The mutual consistency among the viewpoints is ensured by the architecture defined by RM-ODP, and the use of a common object model that binds them all together. More specifically, the RM-ODP framework provides five generic and complementary viewpoints on the system and its environment:

- The *enterprise viewpoint*, which describes the purpose, business requirements, scope and how to meet them through policies for the system.
- The *information viewpoint*, which describes the semantics of the information managed by the system, the information processing performed and the structure and content type of the supporting data.
- The *computational viewpoint*, which describes the functionality provided by the system and its functional decomposition into objects, which interact at interfaces.
- The *engineering viewpoint*, which describes the mechanisms and functions required to support distributed interactions between objects in the system to manage the information and provide the functionality.
- The *technology viewpoint*, which describes the technologies chosen to provide the processing, functionality and presentation of information.

The RM-ODP has consciously been defined in an abstract notation- and representation-neutral language to increase the use and flexibility of the model. However, the lack of precise notations to be used in the individual viewpoints makes it more difficult to develop tools for modeling the viewpoint specifications, the formal analysis of the specifications produced, and the possible derivation of implementations from the system specifications. In order to address these issues, UML4ODP was developed, expressing the viewpoint specifications of RM-ODP using the Unified Modeling Language (UML) to define a UML Profile for each viewpoint language and one to express the correspondences between viewpoints.

Using a the ODP reference model could, in principle, save much effort by successfully partitioning and coordinating design and construction tasks to avoid duplication and gaps, and to ensure the
process of assembly works smoothly with the parts working well together. There are three conditions:

- Third parties providing and using the database and workflow system, have to engage, describing their requirements and conforming to agreements cast in the reference model.
- A sufficient proportion of the software engineers need to engage: Using the reference model when they have questions, and improving it when they find the current answers insufficient.

- Enough of the context in which software engineers are working has to be described at the level at which they work, i.e., at least information, engineering and technology viewpoints.

A wholehearted commitment is needed to reach the thresholds where its benefits are felt by all of those planning, designing, building and maintaining the e-Infrastructure. It is an open question whether this can be achieved with the current state of resources, also for BONSAI.

For workflow sharing and deployment, there are several open source data analytics software frameworks and tools, including:

• Jupyter Notebook: A web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Common uses include data cleaning and transformation, numerical simulation, statistical modeling, and machine learning.

• gCube: A software toolkit used for building and operating Hybrid Data Infrastructures enabling the dynamic deployment of Virtual Research Environments.

• Apache Spark: A general-purpose cluster-computing engine which is very fast and reliable. It provides high-level APIs in Java, Scala, Python and R, and an optimised engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

• Apache Mahout: A platform offering a set of machine-learning algorithms (including collaborative filtering, classification, clustering) designed to be scalable and robust.

• Elasticsearch is a open source search/query engine based on Lucene with an HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java. Official clients are available in Python and many other languages. Elasticsearch is developed alongside a data-collection and log-parsing engine called Logstash, and an analytics and visualisation platform called Kibana. P.S: One commercial vendor of elasticsearch is named BONSAI (https://bonsai.io/).

• OpenLink Virtuoso is the open source edition of the Virtuoso Universal Server, a database engine hybrid that combines the functionality of a traditional relational database management system object-relational database, virtual database, RDF, XML, free-text, web application server and file server functionality in a single system. Rather than have dedicated servers for each of the aforementioned functionality realms, Virtuoso is a "universal server"; it enables a single multithreaded server process that implements multiple protocols.

Provenance

Provenance is “information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness” (W3C 2013). It also covers the concepts of attribution and versioning. The current basis for discussions on provenance is the PROV specification of the World Wide Web Consortium (W3C). The WEST workflow ecosystem specification includes several relevant extensions of PROV. PROV is also supported by workflow management systems like Dispel4Py and Apache Taverna. Komadu is an open source standalone PROV-compatible provenance capture and visualization system.

A provenance management system is also important for systematic issue tracking.

Optimisation

Optimisation is about removing unnecessary constraints for data access, interpretation and processing. The ideal is to have access to *all* the relevant data (including metadata) for a research question, through a single query that specifies both data and algorithms and within an acceptable time results in a human interpretable answer presentation that can be applied either for practical decision making or for improvements to the query, the data, the algorithms, or the presentation format. Reaching the ideal situation would imply an all-inclusive data portal with an automated license-resolver, a mechanism for efficient allocation of processing resources, and a global issue tracker for the improvement part.

Continuous optimisation can be based on tracking of usage patterns to identify bottlenecks. A provenance management service could provide such information.

Semantic linking

Central among semantic web technologies is the Resource Description Framework (RDF) that has come to be used as a generic means to describe information implanted in web resources. Building upon RDF, the Web Ontology Language (OWL) is a knowledge representation language used to describe ontologies, and is a significant factor in many semantic infrastructure modelling projects.