HERVÉ CAUMONT
TERRADUE SRL.

MANAGER, PLATFORM OPERATIONS

PLEASE DESCRIBE THE ROLE OF RESEARCH DATA WITHIN YOUR COMPANY ACTIVITY

First of all, open data is a very important resource for us at Terradue. Our business model acknowledges open data as a business driver. Then scientific data is the main type of open data that our users work with. Terradue develops in particular tools and services to exploit sensor data, such as Earth observation data products. We provide Platform services to a wide range of users having an interest for open research data. These users can be researchers, PhD students, data managers, software developers, system integrators and decision makers. Terradue Cloud Platform provides them with services and tools to connect algorithms on data sources, and to collaborate within virtual communities on the Platform, so they can create new solutions to process open data, extract information out of it, and deliver added value to their partners or stakeholders. Scientific research has been a very active promoter of open data over the past decade. Also the European Commission, together with national space agencies, are developing scientific earth observation satellite programs that acknowledge the principles of open data, applied to developments funded with citizen's money. We have built our Cloud Platform as a collaborative workplace (a.k.a. ellip), for value co-creation out of such kind of research data.

WHAT TOOLS EXIST FOR MANAGING, CURATING, PUBLISHING, DISCOVERING, WRANGLING, PROFILING, MODELLING AND/OR REPORTING DATA THAT ARE AVAILABLE TO YOUR COMPANY?

Terradue is a contributing member of the Open Geospatial Consortium. We contribute as editor of some open standards for data management, for example the OpenSearch Geo & Time extension for the OGC Catalog Service for the Web, or 'CS-W' interface standard. Of course we also implement these standards, as software interfaces, in our Platform services. Interoperability enablers are very important for us, because we acknowledge partnerships and capacity building in user communities as a way to grow and sustain our business. We've actually implemented or integrated quite a lot of such interoperability standards in our Platform's software baseline, covering Cloud Computing APIs and tools, data discovery & access Web Services, and of course data and metadata formats readers. For the later, this type of software is allowing the creation or update of catalog records each time we connect a new data repository to the Platform (most of the time, they are distributed repositories curated by partner organizations), and are supporting our Platform users for the ingestion of data products into their data processing chains and models. The publication of data processing results is making use of tools such as Geoserver (for the visualization as map overlays, within a web browser) or OpenDAP (more suited for data slicing according to some dimensions of interest) and these tools are natively supporting means to catalog the resulting information products or services. We are also

working with technology partners to enhance the products descriptions with more contextual information (information on the data products lineage, on the related products…) by using Linked Open Data (LOD) technologies in order to semantically enrich these descriptions, and make them available over the Web (i.e. contributing data to the Semantic Web).

HOW IS DOCUMENTATION MANAGED? IS THE STORAGE AND CURATION OF INTERMEDIATE DATA CONSIDERED? AND THE DEAD-END WAYS DISCARDED?

Documentation of research data in the domain of Earth observations from satellites is now essentially managed as a digital resource, maintained online, and available through Application Programming Interfaces (APIs). Standards for the description of data products have been matured over the past decade for that domain. Terradue has been an active contributor to it as part of the Open Geospatial Consortium (OGC) but also the Committee on Earth Observation Satellites (CEOS) and the Group on Earth Observations (GEO). Data documentation is a huge effort, also in our domain, but a large corpus is already accessible online. Further work is on-going to make it more systematically machine processable (e.g. via online catalog services for data products discovery and access). As an OGC member, we've worked and still work as a standard editor, especially for the link to user requirements (discovery, access, processing, distribution) and for the promotion of earth observations as a useful source of information in domains such as environment protection studies, water resources management, geohazards monitoring, ocean and coastal research, or Arctic regions modeling, to cite a few. Earth observation satellites provide with reliable and repeatable measurements over time, and have generated decadal time-series fully accessible from online archives (sustained through the long term data preservation programs of the space agencies). They are operated to ensure a global coverage, and they deliver interoperable data products that can be combined within algorithms or models. The policies to manage intermediate products are based on the technological maturity level of the stakeholders, and some best practices answering questions such as: can these products be re-generated on demand at low cost? Can user communities take charge of the curation of some intermediate products? Are the metadata conventions used to document the products well shared across the community?

HOW IMPORTANT IS THE SCALABILITY OF DATA RESOURCES AND DATA-RELATED WORK IN YOUR COMPANY?

Scalability is very important as our customers are dealing with high volumes of data. We cannot project growth models for our business if there is no such scalability. We've been investing a lot in the past to guarantee that our Platform services deliver scalable data processing resources to our users. But also we see now more public funding going into the provisioning of scalable Cloud infrastructures to host the large volumes of Open data generated from the Earth observation programs, and from which our Platform users can also benefit.

WHAT ARE THE "RESULTS" OR PRODUCTS FROM RESEARCH DATA? WHAT HAPPENS TO THESE IN THE LONGER TERM?

We believe that the products from research data should be systematically included as part of a 'value-creation wheel' that the digital economy now allows for. Open and interoperable Cloud services such as Zenodo (an outcome of a past EC PF7 funded project, and now sustained and operated by CERN) automatically provide Digital Object Identifiers (DOI) to research outcomes that are uploaded and registered there, and new services can interface to exploit these results. We've done it for the ESA-funded Geohazards Exploitation Platform, and it works quite efficiently.

WHAT KINDS OF DATA SOURCES AND FORMATS DO YOU WORK WITH? ARE THE DATA SOURCES YOU WORK WITH TYPICALLY OPEN / PUBLIC?

We work with sensor data sources that are curated by third party organizations, and with which we collaborate through consortium agreements or partner agreements in order to connect repositories. We talk of distributed repositories as the connectivity is performed via network bandwidth and Internet protocols. Most of the time the data sources are public, under open data licenses, but we do have some examples of connected data products having usage restrictions. These restrictions can also be applied to open data in some cases (open access for non-commercial use only). These are interesting cases because it means very often that the data provider is also 'open' to discuss terms and conditions for a commercial use! In recent cases, we also had to deal with some level of secrecy and protect the data all along the value-adding chain, because a partner on the Platform is preparing a commercial service. There is still progress to be made in the domain of open data licenses (leveraging the creative commons licensing scheme) but we see practitioners are more and more savvy.

WHAT HAPPENS WITH STANDARDS?

They evolve! So it is sometimes an issue, as the technology and the software implementations sometimes follow these evolutions with disparate timeframes.

WHAT ARE THE MOST DIFFICULT DATA MANAGEMENT REQUIREMENTS FOR YOUR COMPANY? WHAT ARE THE RECURRING BOTTLENECKS?

For us the stake is on the automation of our data management processes. We have Platform services that are operated on principles like data caching or data mirroring that require a lot of attention and careful design, as they impact the Platform's operations costs.

HOW DOES YOUR ORGANISATION DECIDE ABOUT THE CORRECTNESS AND AUTHENTICITY OF DATA FOR ITS RE-USE? FOR INSTANCE, DO YOU TAKE VALIDATION MEASURES OR USE PEER REVIEW?

We partner with data providers, so we do not overcome responsibilities or neither acquire it through buying their data. Ellip, as a collaborative workplace, is more of a matchmaker to ease the way data providers, software developers, system integrators and community managers can work together to innovate and co-create value-added services. Data providers in the Earth observation domain already have internal data management rules and comply to high quality standards. They are trusted. Some data providers having less funding, let's say the initiatives constituting the long tail of research data generation, can still have data holding of interest but less quality control. Our role with regard to correctness and authenticity is to make sure that there is a positive feedback loop to all providers and that our Platform services convey the proper level of information to the potential users of such research data.

CAN YOU ENVISAGE ANY BETTER WAYS TO IMPROVE THE QUALITY OF DATA? DO THESE REQUIRE COOPERATION FROM OTHER STAKEHOLDERS?

Quality assurance and quality control have to be made easier and less expensive. Technology shall help. Efforts are still needed on the technology providers side to help in this goal. Ownership and collaboration remain very important. People working together better understand limits and constraints when reusing data.

DOES YOUR COMPANY MAKE USE OF RELEVANT RDA RECOMMENDATIONS OR EU TECHNICAL SPECIFICATIONS?

Yes, we've made use of: "LEGAL INTEROPERABILITY OF RESEARCH DATA: PRINCIPLES AND IMPLEMENTATION GUIDELINES", by RDA-CODATA Legal Interoperability Interest Group. We adopted it as quality check guidelines, as part of our implementation of procedures for the Ellip platform operations at Terradue.

THERE IS A GROWING AVAILABILITY OF AND DEMAND FOR DATA → DOES THIS CREATE RISKS? VALUE? BOTH?

There is a projected annual growth of data in our sector of 14% up to 2020 (source: Global Satellite-based Earth Observation Market Research Report 2016-2020). This creates opportunities for the creation of innovative services.

DATA INTEGRATION FROM DIFFERENT SOURCES CAN BE QUITE EXPENSIVE, IS THIS BUSINESS MODEL SUSTAINABLE? IF INDUSTRIES ADOPTED GLOBAL INTEROPERABILITY STANDARDS, WHAT WOULD THE GAINS BE COMPARED TO THE LOSSES?

The past EC Framework Programme 7 (FP7) and currently the Horizon 2020 programme funding provide an important vehicle for capacity building, within different types of user

communities. It is also important for the analysis of potential future markets, and for the elaboration of a go-to-market strategy. A nice example was the EC FP7 MELODIES project. The MELODIES project brought together sixteen partners from eight European countries, in order to innovate and develop new services based upon Open Data, combining Earth Observation data with other data sources, and producing new information for the benefit of users. The project could address end-users such as scientists, industry, government decision-makers, public service providers and in some cases even citizens by making use of 'citizen science' approach and techniques.

OPEN DATA FORUM OR OPEN DATA MARKET? IF ONE OF THESE EXISTED, HOW WOULD IT BE STRUCTURED? WHICH ROLES WOULD BE NEEDED? WHO WOULD FULFILL THOSE ROLES? HOW WOULD THE ACADEMIC WORLD AND COMMUNITY DRIVEN INITIATIVES SUCH AS RDA COLLABORATE EFFECTIVELY WITH INDUSTRY?

There are many "Open Data" initiatives. A Forum could provide reviews and success stories. A so called "Open Data market" is quite difficult to qualify. What we see is a growing market for "Custom Information products" or "geospatial intelligence products" making effective use of Open Data as a component or as an enabler of the business model.

HOW DO DATA CREATION, ORGANISATION AND MANAGEMENT PRACTICES NEED TO BE CHANGED TO MAKE DATA INTENSIVE PROJECTS MORE EFFICIENT AND THUS ENABLE RE-USE EFFECTIVELY? COULD YOU GIVE US SOME EXAMPLES OF SUCCESS STORIES REGARDING DATA MANAGEMENT?

We are working at Terradue on two complementary approaches as part of the Ellip vision:

- Automation of tedious and repetitive data manipulation operations (selection, ingestion, publication)
- Creation of collaborative tools allowing added-value activities to occur within data management workflows shared between specialist organisations, each providing its expertise at some point of the workflow (curation, wrangling, modelling)