AAFC-AAC

# Agricultural Data Interest Group and Research Data Alliance

Travel Report to Paris, September 2015

Nicolas Tremblay, agr., PhD, Research Scientist
09/11/2015

## TABLE OF CONTENTS

## INTRODUCTION

This report is a summary of lessons learned and information gathered while attending the workshop of the Interest Group on Agriculture Data (IGAD) of the Research Data Alliance (RDA). The IGAD pre-meeting was held at INRA building, 147 Rue de l'Université from 21 to 22 of September 2015. The RDA 6th Plenary Meeting was hosted in the Conservatoire national des arts et métiers (Cnam), Paris from 23-25 September 2015.

At some place the mention "Re [NT notes]" refers to Annex 2 where notes taken are to be found.

## WHY IT IS IMPORTANT TO ACHIEVE DATA SHARING

- Data management, sharing and analysis play an increasing role in research
- Analysis of historical data (and Big Data as well) helps finding which theory is working in practice (prescriptive analytics suggests decisions)
- It is an opportunity to strengthen linkages between research, development and policy in a continuum: (Big) Data → Information → Knowledge → Wisdom (policy makers, industry, societal stakeholders)
- Open data is a necessary step in the Big Data vision where there is one massive linked data pool across disciplines and strong computational capabilities
- This site summarizes why open data is important: http://landportal.info/blog-post/2015/07/will-you-share-data
- A better use of agronomic data would make it possible to provide recommendations for management practices for sustainable use of soil resources and improvement of agricultural production
- The goal is to achieve spatially explicit and evidence based recommendations

## ISSUES IN DATA SHARING

- Scientists sit on their data until they retire and then it is lost (Seishi Ninomiya, Tokyo, Japan)
- Why not deposing data and metadata? Reasons are 5% technical and 95% cultural
- Currently, there is little or no incentive for scientists to share their data other than good will, which is often limited by the lack of time
- There is a need for recognition in data sharing. If the data is published and, better of, reused by someone else (therefore accessible), this should be highly valued in the annual appraisal and the record for promotion purposes
- It is not enough to put~~t~~ data in archives (data dumps). Data in a dump is impossible to find by others. It is admitted that discoverability, however, is difficult to achieve as there are fundamental methodological issues for reproductibility and transparency
- There are currently no metadata standards. Some work is in progress in France on ontology: Élisabeth Arnaud CGIAR Consortium office, France
- The new area of smart farming (monitoring, planning & control, Crop/animal level, farm level, market level, environmental level) is particularly challenging as it is hard to combine data in such context

## DATA MANAGEMENT PLANS

- The goal is to achieve data circulation; used and reused
- Partner organization create their own database and make it accessible through the internet
- The challenge is to link databases worldwide such that integrated search is possible. Indeed, researchers from outside agronomy will like to access this data as well. Agriculture is not a silo
- Data management plans are a mean to make the data available and interoperable (access and aggregate data of diverse source, and be able to combine it and use in information systems different than those of origin). Data management plan should be figured out right from the start of a research project
- For a scientist, however, the true requirements are "impossible to meet". More importantly than technological capacity, support has to be in place in the form of a cohort of data professionnals
- Librarians are data management specialists. If they are involved at the beginning of a data management plan the effort is 1/5 of what it would be afterwards
- The Canadian Association of Research Libraries (CARL) launched a project to develop a library-based research data management network in Canada, Portage
- http://www.carl-abrc.ca/uploads/SCC/Portage-External-2-Dec-22-2014.pdf
- Agreements between actors are needed in the data chain

## RESEARCH DATA ALLIANCE (RDA)

- Every group of discipline thinks that its situation is unique while, actually, different disciplines have a lot of communalities
- RDA is a forum for exchange, with no legal entity, that acts as a catalyst, but does not provide funding
- Its organisational structure can be found at : https://rd-alliance.org/organisation.html
- RDA works from use cases conveyed by Interest groups (with deliverables) e.g. Agricultural Data Interest Group (IGAD) https://rd-alliance.org/groups/agriculture-data-interest-group-igad.html
- Interest groups may have working groups under them. Working groups never stand alone, the are always under an interest group
- e.g. Wheat Data Interoperability Working Group
    - "The Wheat Data Interoperability Working Group aims to provide a common framework for describing, representing linking and publishing Wheat data with respect to open standards.Such a framework will promote and sustain Wheat data sharing, reusability and operability. Specifying the Wheat linked data framework will come with many questions: which (minimal) metadata to describe which type of data? Which vocabularies/ontologies/formats? Which good practices? Mainly based on the the needs of the Wheat initiatiative Information System (WheatIS) in terms of functionalities and data types, the working group will identify relevant use cases in order to produce a "cookbook" on how to produce "wheat data" that are easily shareable, reusable and interoperable."
    - Deliverables: Wheat data guidelines, Wheat data vocabularies portal
- Membership Benefits
    - Being seen to act as early adopters for newly developed standards and protocols

    - Adding influence to their work on data interoperability in their sectors, markets and geographies
    - Having a voice inside RDA, providing advice on the needs of their sectors and the problems faced in data exchange
    - Providing advice to the RDA Council through the Organisational Advisory Board
    - Being recognised on the RDA Website and at RDA Meetings as a leader in world data sharing and interoperability
    - Receiving regular briefings on the progress of RDA's work
    - Be able to participate in the 6-monthly Organisational Assembly that takes place at RDA Plenaries interacting with other Organisational Members (OMs)
- Membership costs
    - < than 50 employees $1000 US per annum
    - 50 and <250 employees $2000 US per annum
    - >250 employees $10,000 US per annum
- Research Data Canada (RDC) http://www.rdc-drc.ca/
    - "RDC is a forum in which multiple stakeholders come together to build research data stewardship in Canada. RDC is a collaborative effort to address the challenges and issues surrounding the access and preservation of data arising from Canadian research. The challenges surrounding research data access and preservation must be addressed collectively from a national perspective with the participation of all parts of the research community, including researchers themselves who create and use this data. The membership of RDC comprises a wide variety of individuals and stakeholder organizations, all of whom have an interest in and role to play in ensuring that the infrastructure, processes, and support are in place to ensure research data in Canada is both accessible and preserved. Members include organizations in broad-band networking, research funding agencies, universities, NGOs, government research organizations, and research libraries."
    - Walter Stewart is the co-ordinator https://www.rd-alliance.org/about/organization/key-profiles/walter-stewart.html
    - Agriculture and Agri-Food Canada is not currently part of RDC. However, the following departments and agencies are the Canadian Space Agency, Environment Canada, National Research Centre, Treasury Board Secretariat, Industry Canada and Statistics Canada http://www.rdc-drc.ca/about-us/partners/
    - In the US, the National Science Foundation (NSF) is supportive of RDA

## PARALLEL INITIATIVES TO RDA

### HORIZON 2020

- The EU Framework Programme for Research and Innovation
- Horizon 2020 draft work programme 2016-2017: Reingineering of an entire production sector around data management

## BIG DATA EUROPE PROJECT

- Mobilising the data revolution is the major priority for international science
- http://www.big-data-europe.eu/food/
- BigDataEurope workshop on "Big data for food, agriculture and forestry: opportunities and challenges
- http://www.big-data-europe.eu/bigdataeurope-workshop-on-big-data-for-food-agriculture-and-forestry-opportunities-and-challenges-held-in-paris-on-22-september-2015-the-main-challenge-is-variety/
- V's of big data (Volume, Variety, Velocity, Veracity) are also typically used in agricultural, food and environmental research where we are not talking about an extremely large Volume; other domains have much more voluminous data. It is not that they come with a high Velocity, especially compared to other domains. In many cases, their Veracity is quite high. But in agriculture, data Variety matters: you need to combine multiple, heterogeneous data types and formats from several sources, trying to solve the information problems and support decision making of the relevant stakeholders.
- http://www.ibmbigdatahub.com/infographic/four-vs-big-data

## ELIXIR

- ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.
- https://www.elixir-europe.org/

## W3C

- The World Wide Web Consortium (W3C) is an international community that develops open standards to ensure the long-term growth of the Web.
- http://www.w3.org/Consortium/mission.html

## GODAN

- Global Open Data for Agriculture and Nutrition (GODAN) works to support global efforts to make agricultural and nutritionally relevant data available, accessible, and usable for application worldwide
- http://www.godan.info/
- GODAN promotes collaboration to harness the growing volume of data generated by new technologies to solve long-standing problems and to benefit farmers and the health of consumers.
- Partners are: the US Government, the UK's DFID, the Netherlands Government, the Open Data Institute, FAO, CTA, CABI, CGIAR and GFAR.
- GODAN interested in private/public relationships
- USDA is strong in GODAN
- Canada already much involved in GODAN

## CODATA

- CODATA (International Council for Science: Committee on Data for Science and Technology) http://www.codata.org/about-codata
- Priority for CODATA: Data science for research - data analysis skills
- Semantic examples in the "landbook" at landportal.info

## COPO

- Collaborative Open Plant Omics
- Genomics and Computational Bioscience
- “Publication models are moving away from the traditional "data late" approach, and are shifting towards the "data early", with particular pressure being made by funding agencies to see data publicised quickly. Alongside the wealth of these large plant science datasets held in public and private laboratories around the globe, there are a large number of tools to help researchers disseminate, analyse and publish those datasets. However, the disparate nature of the tools, data formats and scientific problems in light of this expansive experimental development has resulted in a lack of interoperable, production-quality software available for data analysis and dissemination.”
- “COPO addresses this disparity in interoperability and easy access to important services in the 'omics data realm. We will develop a framework to utilise existing services to facilitate the description, deposition and publication of datasets, but also to enable the identification and citation of datasets, thereby increasing transparency and reproducibilty. Promoting reward for making data available is a central aim of the project.”
- https://documentation.tgac.ac.uk/display/COPO/Overview
- https://documentation.tgac.ac.uk/download/attachments/6357358/davey_EPSO-COPO-overview-20140623.pdf?api=v2

## AGRICULTURAL INFORMATION MANAGEMENT STANDARDS (AIMS)

- “AIMS is a portal with information about and access to standards, technology and good practices. It is also a forum for connecting information management workers worldwide and for discussing open access and open data. AIMS stands for collaboration and interoperability”
- http://aims.fao.org/

## CANARIE

- “CANARIE designs and delivers digital infrastructure, and drives its adoption for Canada’s research, education and innovation communities. CANARIE keeps Canada at the forefront of digital research and innovation, fundamental to a vibrant digital economy.”
- http://www.canarie.ca/
- https://rd-alliance.org/about/organization/key-profiles/nancy-carter.html

## EXAMPLES FROM OTHER ORGANIZATIONS

### INRA

- INRA has gone through a process that is an inspiration for other research organizations
- It is the result of a process ranging from 2009 to 2012, ending with a report of the INRA scientific council:
- 2012 Report of INRA scientific council + data management and sharing: recommendations
- http://prodinra.inra.fr/ft?id=%7bE261B98C-86C6-42D8-AD89-622FAB942C55%7d
- Key issues and actions identified by INRA:
- Pain points (for Experimentation, observation and simulation scientists)
    - Data standardization
    - Exchange, transfer
    - Capturing metadata automatically
    - Gap in semantic coverage
    - Metadata may be strategic (e.g. protocols, methods)
    - Some metada or data are sensitive (geographic information about epidemiologic cata or GMO data)
    - Exchange of large datasets
- Scientists wish lists
    - Recognition of the contributors
    - Recognition of data sharing as first class skil at the institutional level
    - Data papers/training
    - A data portal with access to all the data shared by INRA
    - Harmonization of metadata standards and vocabularies
- Digital repository in construction
    - Based on the IT environment provided by the IT department
    - Covers both active and historical data
    - Outsource the long term preservation
    - Leverage the many existing data repositories
    - Conforms to the Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model ISO 14721:2012. “An OAIS is an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. It meets a set of such responsibilities as defined in this International Standard, and this allows an OAIS archive to be distinguished from other uses of the term "archive". The term "open" in OAIS is used to imply that ISO 14721:2012, as well as future related International Standards, are developed in open forums, and it does not imply that access to the archive is unrestricted.”
    - http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284
- See also: datacite.inist.fr/IMG/pptx/datacite_inra_2_1.pptx

### WAGENINGEN UNIVERSITY

- Re [NT notes]: Data management plans: means and goals? (Hugo Besemer, Wageningen)

- There is an obligation to make data available (cultural change)
- Data management plans are mandatory PhD projects and research groups since April 2014
- A data management planning course is available for PhD candidates since 2012
- Important to consider benefits for the scientists for example in the form of data citations (way to be recognised for contributions)
- Re [NT notes]: CODATA, Open Science Policies and Capacity Building (Sidmon Hodson, Committee on Data for Science and Technology, CODATA, France)

## ANNEX 1: TECHNICAL ELEMENTS OF DATA SHARING

There are elements to put together to achieve adequate data sharing but no tool is fully compliant with all needs

### VOCABULARIES / THESAURI

- It is important to have reference taxonomy to enable the semantic interoperability that will facilitate the content discovery.
- "AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc. It is published by FAO and edited by a community of experts. AGROVOC consists of over 32,000 concepts available in 23 languages. You can use AGROVOC to look up the common name of a plant in a language that you do not master, or to find relations between a commodity and the crop from which it is produced. You can use it from inside your content management system (e.g., Drupal) to organize your documents or web site. You can also use AGROVOC as a hub to access many other vocabularies available on the web." http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
- National Agricultural Library Agricultural Thesaurus http://agclass.nal.usda.gov/
- The CAB Thesaurus has many potential uses by individuals and organizations indexing their own information resources for both internal use and on the Internet. http://www.cabi.org/cabthesaurus/
- Other examples: DCAT vocabulary, VOID vocabulary, SDMX - datacube vocabulary
- Issues: Missing vocabularies for protocols, formats. Some missing properties in existing vocabularies

### SEMANTICS

- Semantic interoperability facilitates building of data services that reuse and combine data from different sources
- "Massive amount of data, in many languages, is being produced in agriculture and related disciplines. The advantage of having large amounts of data is now largely recognized, but it goes hand-in-hand with the need to "understanding" it - to understand what it is about and how it relates to other pieces of data. For example, what are the local varieties of a crop, more resistant to climate change than the commercial ones? What information is available about them and their cultural and environmental conditions? Would we be able to find them across the different names used in their region of distribution? And, can we enable small institutions as well as the

big ones to publish and organize their data so as to be compatible and integrated with the data produced by other entities? Semantics is the mechanism that allows us to address problems like these."

- http://aims.fao.org/agrisemantics-workshop-2015
- "In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS."

## STANDARDIZATION

- Interoperability exists only if the data is operable in other systems than the one it has been created in.
- Issue: Lack of interoperability metadata in existing tools
- The Statistical Data and Metadata eXchange (SDMX) is an initiative to foster standards for the exchange of statistical information.
- https://en.wikipedia.org/wiki/SDMX
- Geographic information metadata: ISO 19115-1:2014 defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services.
- http://www.iso.org/iso/catalogue_detail.htm?csnumber=53798

## ONTOLOGY

- Ontology deals with questions on how entities may be grouped, related within a hierarchy, and subdivided according to similarities and differences.
- AgroPortal is a proposition for ontology-based services in the agronomic domain https://hal.inria.fr/IBC/lirmm-01172232v1
- The Plant Ontology describes plant anatomy and morphology and stages of development for all plants. The goal of the PO is to establish a semantic framework for meaningful cross-species queries across gene expression and phenotype data sets from plant genomics and genetics experiments http://www.cropontology.org/

## ANNEX 2: NICOLAS TREMBLAY'S NOTES

### AGRICULTURAL DATA INTEREST GROUP (IGAD) PRE-MEETING

**INRA building, 147 Rue de l'Université**

**Research Data Alliance**

Participants: about 50

Introduction (Odile Hologne, INRA) – History of the Agricultural Data Interest Group (IGAD)

- Goteborg, less than 10 people to start the initiative
- Today, more than 50
- INRA, 2nd after USDA-ARS, in the
- 8000, with 2000 researchers, 17 research centers
- Department of scientific information
    - Librarians, information technology

Devika Madalli (Bangalore)

- Focus on soil databases

- UN-FAO
- Now 24 countries
- Wheat data interoperability group

Imma Subirat (FAO)

- Pre-meeting for the plenary
- Showcasing what is going on in the agricultural community
- San Diego: request to defining better the lines of actions
    - Institutionnal issues
    - Increase access availability
    - Interoperability
    - Another thematic group: cereal, proposed to RDA
- Increase participation
- Wednesday at RDA plenary (panel)
- Thursday, together with wheat data interoperability working group
    - Status, deliverables
- Big Data Europe project

**Data management plans: means and goals? (Hugo Besemer, Wageningen)**

Slides available at: www.slideshare.net/HugoBesemer

http://fr.slideshare.net/hugobesemer/data-management-planning-52923625

- Needed 2 years of discussion, on going. Social sciences, particularly touchy
- Data management planning course for PhD candidates since 2012
- DM plans mandatory for PhD projects and research groups since April 2014
- Policy supported by:
    - Support hub for all questions
    - Facilities for data publishing
    - Code repository
- Still working on:
    - Storage "archiving"
    - Electronic laboratory notebooks
    - Guidelines for ownership
- Institutional data management planning
    - Who benefits?
        - Policy makers: there is a policy
        - PhD researchers
        - Research groups
    - What is needed
        - People and organizational dynamics
        - Develop a common language!
- Two cultures (according to C.P. Snow)
    - Lord Snow
        - Literary intellectuals
        - Natural scientists
    - Many aginfo meetings
        - Knowledge managers
        - Techies
    - What I am experiencing now
        - Infrastructure builders
        - Empirical scientists
- Different meanings: Data
    - Infrastructure builders
        - Data is the evidence that supports "truth"
    - Empirical scientists
        - Data resides somewhere and comes in a specific "physical" format
- Metadata
    - Infra
        - Data about data
    - Empirical sci
        - Anything goes: templates, parameters used, data models, laboratory notes, annotations
- Data management roles
    - Infra
        - Scientists sit on their data and should be convinced to depose it
    - Empirical sci
        - Produced in chains under informal agreements
- Documentation
    - Infra
        - Documenting a static dataset at project file and parameter level
    - Empir
        - May include laboratory notes etc during research
- Storage, archiving and data publishing
    - Infra

        - Fluid terminology
    - Empir
        - Storage is for daily use during research
        - Archiving can be for data at later stage
        - People like "data publishing" for datasets deposited in repositories
- Legal issues
    - Infra
        - Open licenses for data re-use
    - Empir
        - Agreements between actorsin the data chain
        - Interested in open licenses, also as consumers
- What about IGAD?
    - Institutional issues
        - If we want to address institutional issues we need to be aware of different cultures and languages
        - I did not mention anything specifically agricultural, but that does not imply that there are no specific institutional issues
        - For us, things started with learning and training activities
- Putting in archives is not enough. Hard to find by others
- www.slideshare.net/HugoBesemer

**CODATA, Open Science Policies and Capacity Building (Sidmon Hodson, Committee on Data for Science and Technology, CODATA, France)**

- Data revolution: challenges and opportunities
    - The digital age has brought a data revolution that presents science with major challenges and opportunities
    - Opportunities
        - Can gather unprecedented volumes and types of data and analyse them more quickly
    - Challenges
        - Data infrastructure, networks and analysis
        - Fundamental methodological issues for reproductibility and transparency
        - For science systems, technical and human
    - Data for research should be intelligently open: accessible, intelligible, useable
    - Creating a world that counts: mobilising the data revolution
- Vision
    - Available online, open, all publications open and on line and for them to interoperate
        - Obligation to make data available; cultural change / have to consider "self promotion"
        - Data citation: way to be recognised for contributions
    - Raw data -> derived and recombined data -> literature
- Availability is not enough; interoperable too
- CODATA strategy
    - Mobilising the data revolution
    - Exploiting the data revolution is **the** major priority for international science
    - Deliver benefits
    - Part of "Science International" and Open Science Capacity Building Initiative
    - Leading the discussion on "International Accord on Open Data and Open Science"
    - Will launch a broader international Open Science Capacity Building Initiative

- Open data + Big Data analytics Ecosystem
    - Opportunity
    - To produce good science
    - Good science - open data forum and policies - infrastructure - training...
- Data policies
    - FORCE11 is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. Individually and collectively, we aim to bring about a change in modern scholarly communications through the effective use of information technology. Scholarly communication by means of semantically enhanced media-rich digital publishing is likely to have a greater impact than communication in traditional print media or electronic facsimiles of printed works. However, to date, online versions of 'scholarly outputs' have tended to replicate print forms, rather than exploit the additional functionalities afforded by the digital terrain. We believe that digital publishing of enhanced papers will enable more effective scholarly communication, which will also broaden to include, for example, the publication of software tools, and research communication by means of social media channels.
    - https://www.force11.org/group/data-citation-implementation-pilot-dcip
        - From principles to practice
        - Journals, institutions have to be involved
    - http://bit.ly/data_citation_principles
    - FINAL Joint Declaration of Data Citation Principles and Endorsement page
        - https://www.force11.org/group/joint-declaration-data-citation-principles-final
- Training
    - Capacity building
    - Summer schools
- Questions
    - Private sector involvement?
    - Microsoft research, Dell, etc. involved in courses
    - Priority for CODATA: Data science for research - data analysis skills

**Agricultural Soils Research Data in Tasmania, Australia; Policies, collaboration and sharing (Darren Kidd, Tasmania)** (remote presentation)

- Geospatial data
- Land use, soil condition, vulnerable soils, salinity, sodicity...
- Providing spatial data for use in agricultural or environmental research (modelling)
- Benefit of data storage, prescribed format, enhance sharing
- Publicly available, can make comments, open
- GeoNetwork (open source)
- Standard properties for consistency
- Conforms and contribute to the GlobalSoilMap
    - Specifications respected
- 18,000 geo-referenced soil profiles have been located, and are being harmonised and digitized. Present DB holds 6500 sites.

**Soil Data Availability in support of Agriculture Development and Environmental Protection (Pandi Zdruli, CIHEAM, Bari, Italy)**

- Threats affecting European soils
    - Soil Thematic Strategy COM(2006)231 final
- Collection of soil data
    - Soil surveys
    - Expensive: 3 k$, including lab analysis
    - How to organize and display soil data?
- Jeju Korea
    - Common language for description
- European Soil Portal - Soil Data and Information Systems
    - Some freely available, some restricted due to copyright rules, legal implications
    - LUCAS will establish a long term soil quality monitoring system in Europe
- INSPIRE compliant metadata profile for soil geographic datasets. Metadata editor
- DIGISOIL concept of data storing and processing
- Soil data information in the USA
    - NASIS, STATSGO, SSURGO, MUIR
    - Freely available
- ISRIC www.isric.org
- Africa Soil Information Service
    - 18532 unique soil profile records, almost all georeferenced
- Harmonized World Soil Database (developed by FAO)
    - Interactive map
- Challenges for soil data collection and distribution
    - Global soil partnership (by the FAO)
    - Pillar 4: Enhance quantity and quality of soil data and information, data collection, analysis, validation, reporting...
- Conclusion
    - Soil data should be easily available
    - Web based platforms must be friendly to use
    - Old soil data must be replaced with new ones, especially in Africa and developing countries
    - Remote sensing technology offers great opportunities for quick data collection
    - Yet, field soil surveys are crucial for validation
    - Interaction and integration between international and national institutions on data distribution should be strengthened
- Pedotransfer rules and functions system to approximate soil parameters for particular sites
    - Not always reliable
    - Need to go and sample
- pandi@iamb.it
- Global soil map no longer funded by Bill Gates fundation

**Soil Research Data Policies, Data availability and Access, and the Interoperability challenge. A Data Management and Sharing Plan for Soil Open Data (Giovanni L'Abate, CREA)**

- Soil data type & formats
    - Soil observations
        - GIS point data

    - Soil maps (polygons)
        - Spatial features or printed maps, WebGIS tools (WMS, WFS, WCS)
    - WMS
        - AtomPub, GIF, GeoRSS, JPEG, KML, OpenLayers...
    - WFS
    - Direct download of spatial features
- Standards
    - ISO SoilML, WFS, WMS, and WCS
    - INSPIRE has brought together other standards
    - OGC observation and measurements - XML
        - Simple feature access
    - World reference Base, 2th edition (2006)
    - USDA Soil Taxonomy
- Capture methods
    - Spectral signatures
    - Soil samples
- Ethics and intellectual property
    - Database Directive (1996) protects the producer of a database
    - How to cite
- Data sharing and reuse
    - Italian Open Data License
- SISI webGIS application
    - https://aginfra-sg.ct.infn.it/sisi
    - Observed soil profiles and derived soil profiles
- Deposit and long-term preservation
    - Select data of long-term value
    - Safeguard the data behind the graph
    - Assure that your data will remain accessible
- The interoperability challenge
    - Semantic interop facilitates building of data services that reuse and combine data from different sources
    - GIS servers commonly use the same protocols but DB are still highly customized
    - ESRI white paper on spatial data standards
    - SISI compliance with INSPIRE and IUSS models
- agINFRA Soil Terms Vocabulary
- SQL code to access soil data
    - AGRIS implementation (Germoplasm data)
- http://abp.entecra.it

**The Soil Research for Development platform for sharing data and information on agronomic trials to develop options for integrated soil fertility management in sub-Saharan Africa (E. Jeroen Huising, Nigeria)**

- TSBF platform intervention areas
    - Better use of agronomic data to be able to provide recommendations for management practices for sustainable use of soil resources and improvement of agricultural production
    - Land degradation and soil ecosystem sustainability
    - Integrated management of soil health and fertility
    - Last mile delivery and dissemination

- Strengthen linkages between research, development and policy
- Want to achieve spatially explicit and evidence based recommendations
    - Survey and inventory data
    - Agronomic trial data
        - Currently no metadata standards
        - (some work in progress in France on ontology: Élisabeth Arnaud CGIAR Consortium office, France)
    - Analysis and interpretation
    - ISFM recommendations
- Improve discovery of data
    - Conduct survey / questionnaire
    - Collect information on datasources
    - Allow to post information
    - Gather information on internet
- Improve access and fitness for use
    - Documentation and metadata
- Improve interpretation and synthesis
- Harmonize data collection
    - Fulfills requirements: protocols (design, sampling and measurements for the various types of crops and cropping systems), data collection templates, data quality control and error detection, minimum data sets
- Create data infrastructure
    - Create data repository
    - Partner organization create their own datagbase and make this accessible through the internet
    - Link databases such that integrated search is possible
- Operation of the platform
    - West and Central Africa
    - Build on existing networks
    - Membership free and voluntary
    - Use of TSBF platform website for communication and organiza biannual conferences
    - Hosted by IITA in Nigeria

**Linked open data for land governance (Neil Sorensen, landportal.info)**

- Land portal pioneers linked open data for land governance
- Content fragmented, difficult to locate, hard to reuse
- Metadata not available or organized
- Leading online source for land issues: landportal.info
    - Social
    - Economic
    - Etc
- Semantic: in "landbook" on the portal

**OGC standards for the Interoperability of agriculture models: data and processes at the same level (D. Leibovici, U. Nothingham, UK)**

- Agriculture modelling, geospatial and temporal
- Production models (crop), social, economical, sustainability, climate change
- GRASP: Geospatial resource of agricultural species and pests with integrated workflow modelling
    - part of CropBASE (CFF)
- Spatial data infrastructure
    - GermplasmDB (genotypic variations)
        - Wheat eyespot disease: harvest reduction model,
- Workflow composition of data & processes
- Quality & error propagation
    - Need metadata properly encoded
    - ISO 19115-1 / CSW / metadata brokers
    - Data quality & processing quality
    - Data and process discovery (metadata)
    - H2020 EINFRA-9

**Implementation of a service for research data management and sharing in INRA (Ester Dzalé Yeumo Kaboré, INRA, France)**

- INRA
    - 2009 political awareness of INRA CEO: INRA policy for data management and sharing
    - 2012 Report of INRA scientific council + data management and sharing: recommendations
    - Implementation
- What data?
    - Genomic, observation, experimental, simulation, social, genetic
    - Text, numeric video, audio, images
    - Purchased, big, small, secondary, historical
- Current sharing practices: the point of view of some scientists
    - Genetic & genomic
    - Experimentation, observation and simulation
    - Social sciences
- Pain points (for Experimentation, observation and simulation scientists)
    - Data standardization
    - Exchange, transfer
    - Capturing metadata automatically
    - Gap in semantic coverage
    - Metadata may be strategic (e.g. protocols, methods)
    - Some metada or data are sensitive (geographic information about epidemiologic cata or GMO data)
    - Exchange of large datasets
- Scientists wish lists
    - Recognition of the contributors
    - Recognition of data sharing as first class skil at the institutional level
    - Data papers/training
    - A data portal with access to all the data shared by INRA

    - Harmonization of metadata standards and vocabularies
- Website put in place: "Open Science"
- Digital repository in construction
    - Based on the IT environment provided by the IT department
    - Outsource the long term preservation
    - Leverage the many existing data repositories
    - Conform to the OAIS reference model
    - Cover both active and historical data
- DOI Minting service
- Training
- Questions
    - Recognition of contribution: difficult in meta-analyses

**How to describe a dataset? (Valeria Pesce, GFAR, Italy)**

- Interoperability issues
- Discoverability difficult to achieve
- The "instances" of the dataset "available for access or download in one or more formats" are called "distributions". A dataset can have many distributions (examples of distributions include a downloadable CSV file, on API or on RSS feed)
    - Applications will look for this
- Interoperability: only if operable in other systems, otherwise, not
    - Retrived, processed, re-used and re-package
- What applications need
    - Name, author / owner, date...
    - Metadata
        - Coverage
        - Technical specifications
        - Conditions
        - Dimensions
        - Semantics of the dimensions (units of measure, time, granularity, syntax, reference taxonomies)
- Partial answers in existing vocabularies
    - DCAT vocabulary
    - VOID vocabulary
    - SDMX - datacube vocabulary
- Coverage of the dataset
    - Common Dublin Core properties (DCAT re-uses these DC properties)
    - The values of these properties have to be understood by machines
    - There is no authority vocabulary for types of data
- Conditions for re-use
    - DCAT re-uses the license DC property
    - W3C DCAT > DCAT AP
- Technical properties
    - Specs to retrive and parse a distribution of a dataset (format, protocol, etc)
    - VOID can help with the protocol metadata but only for RDF datasets
- Dimensions and their semantics
    - DCAT does not describe the dimensions of a dataset except for a reference to a standard

- SDMX: data structure and dimensions (Statistical Data and Metadata Exchange)
- Tools for managing dataset metadata
    - CKAN, Dataverse, OpenAIRE, DataCite, Dryad, CIARD RING
- Major outstanding issues
    - Some missing properties in existing vocabularies
    - Missing vocabularies for protocols, formats
    - Need for more standardized semantics for dimensions
    - Lack of interoperability metadata in existing tools
- No tools that are fully compliant with all needs

**From GACS to agrisemantics - steps forwards (Caterina Caracciolo, FAO of the UN, Italy)**

- Global Agriculture Concept Scheme
    - Make a common repository of terminological and conceptual information in agriculture
- Many similar initiatives in the area, with both some overlaps and references, but no coordination
- Thesauri
    - AGROVOC, CAB Thesaurus, NAL Thesaurus
    - Overlaps between these thesaurus
    - Similarity in use
- Demo, 15,000 concepts
    - Organisms, places, chemicals, "general" concepts
    - Information available results from a "merge of the three"
- Ongoing
    - Finalizing the resolution of "lumps"
    - Clean up labels, e.g. to harmonize use of sing/plur
    - Move in a fully functional editing environment
    - Decide on needed concept types e.g. for organisms, chemicals, ...
    - How to deal with non-identical hierarchies when merging mapped concepts
    - Define general policy for dealing with scientific and common names of organisms (are different things? or different names for the same things?)
- Interoperability
    - Goal: access and aggregate data of diverse source, and be able to combine it and use in information systems different than those of origin
    - Given a set of concept schemes, establish correspondences so that data described can be merged correctly
- After GACS - Agrisemantics
    - Goal - a BMGF project
    - Project proposal for Bill and Melinda Gates fundation
    - Expand GSCS in terms of
        - Coverage
        - Typology of resources involved
- http://aims.fao.org
- http://aims.fao.org/agrisemantics-workshop-2015

**An approach for identifying the issues and metrics that define and measure sustainable development in agricultural supply chains (Ruthie Musker, UC Davis, USA)**

- Goal: structured, comprehensive methodology and tools to help stakeholders in the food industry make sustainable decisions along the supply chain
- Needs to be useful in any location
- Methods of measuring sustainability
    - Issues of sustainability (AGROVOC terms)
    - Building an ontology
- Wiki as a platform (Semantic Wiki)
    - Links agricultural data by food system
    - Can show how users are using indicators and how they are categorizing them (language)
- Future plans
    - Add information about resilience in food systems
    - Organize by commodity, location, goal, etc.
    - Partner finder
    - Depends on what users want
- Conclusion
    - How can we visualize all of this information?
    - How to effectively host datasets on the wiki?
    - Would it be appropriate for your datasets?

**Challenges in Normalizing and Disambiguating Organization Names (John A. Fereira, Cornell University, USA)**

- Creating a system of user profiles, their affiliation (Organization) and other related information (Agri-Profiles)
- AgriProfiles based on VIVO software with custom harvesting tools
- ONLD: Organization name Linked Data

**Facilitating data discovery & sharing among agricultural scientific networks: presentation of case studies (Nikos Manouselis, Agro-Know, Greece)**

- nikosm@agroknow.gr
- Agro-Know: captures, organizes and add value to information
- Service "AKstem" for research information
    - Across networks
    - www.akstem.com/gfsp
- Within a collaboration...
    - Many organization
    - Working on common topic
    - Dissemination around the world
    - E.g. Global Food Safety Partnership
- Typical challenges
    - Sharing topic-specific from larger generic data sources
    - Combining information from multiple data sources
    - Providing simple, seamless search services to users

        - e.g. combining information coming in different metadata formats, through both harvesting and querying products
- Tool: Map your network
    - Create an online map of your partners and their activities
- Next challenges
    - Search & results mashing up different information/data types from multiple sources
    - Linking from one information resource to another, regardless of format & type
    - Going beyond metadata, into data searching, processing, visualising, etc.

**Development of interoperable platform for agricultural data exchange and applications in Japan (Seishi Ninomiya, Tokyo, Japan)**

- Policy of Japanese government
    - Agr data standardization
    - Open data (discussion just starting)
        - Scientists sit on their data until they retire and its lost
- W3C Agriculture Community Group (GC)
    - Discussion not much active so far
- Agriculture information service on modeling and scenario simulation
    - Different platform collect data for agriculture
    - Systems are isolated and must become interoperable
    - Standard, common, open API is needed
- Meteorological data needed
    - MetBroker
        - To unify Agro-Climate DB, data from field, GD-DR&TR
        - Application layer: rice model, cassava model, vegetable model
        - Just 15 scientists in Japan have developed application using MetBroker
            - Rice rust disease prediction
    - Unified access method to a weather databases
    - 17 weather DB, 86,000 weather stations
    - Java class library and web interface (XML, JSON, CSV, ...)
    - DB change their structures often, need adjustments
    - Hard to maintain
- Farm management information systems (FMIS)
    - Each FMIS has proprietary data without standard data exchange method
    - Very inefficient use of data and programs
    - A new data expression for farming data is being developed for data exchange: FIX-pms
    - Current range FIX-pms covers is limited to farm work and production process management
    - Structure based on agroXML
- Cloud Open Platform in agriculture (CLOP)
    - Concept
- Interoperable info platform
    - Sensing/data collection layer
    - Information platform
    - Application service layer
    - Web service
    - Users
- Structure of API mashup: 4 layers
    - Term, code layer

    - Data content layer
    - Data format (container) layer
    - API layer
- Framework for model development
- Questions:
    - Soil data? Very heterogenous, difficult to integrate

**Ontology-based services and knowledge management in the agronomic domain (Pierre Larmande, Instute of Research for Development, France)**

- AgroPortal
    - https://hal.inria.fr/IBC/lirmm-01172232v1
    - Better related to agronomy than the BioPortal and hopefully more used
    - To develop and support a reference ontology repository for the agronomic domain
    - Reusing the NCBO bio portal technology
    - Enable use of agronomic related ontologies
    - Contains already 29 ontologies, 40 soon expected
    - Use cases
        - IBC rice genomics, wheat, lovinra, crops
- Agronomic Linked Data (AgroLD)
    - Semantic web based system that integrates data from South Green Bioinformatics node
    - www.agrold.org

**Crop ontology: harmonizing semantics for agricultural field data (Elizabeth Arnaud et al., CGIAR Consortium, France)**

- Problem
    - No naming convention for variables and methods of measurement which are heterogenous
    - Confusion between traits and variables
    - No semantic coherence
    - Definition and measurements are different
- Needed
    - Crop traits
    - Experimental design
    - Environmental factors
- www.cropontology.org
    - 20 crops
- Standard variable
    - Property (trait) + method + scales/units
    - Unique name
- Online vizualization
    - Trait, Methods, Scales & Standard Variables
- Google Cloud & API
    - Crop Ontology Curation Tool
    - 1410 agronomic traits already documented
- Agronomy ontology (work in progress)

    - Align with the International Consortium for Agricultural System Application (ICASA)
    - ICASA Master Variable list
    - 600 standard variables already defined
- Planteome pilot project
- Mapping crop ontology terms across species and to the reference ontologies

**Collaborative Open Plant Omics (COPO) (F. Shaw et al., TGAC, UK)**

- Open source
- https://documentation.tgac.ac.uk/plugins/servlet/mobile#content/view/6357358
- Goals of COPO
    - Enable standards-compliant data collection, curation and integration
    - Enhance access to data analysis and visualization pipelines
    - Facilitate data sharing and publication to promote reuse
- Reproducibility depends on
    - Reducing reinvention
    - Describing methods and data
- Why not deposing data and metadata?
    - Lack of interoperability
    - 5% technical, 95% cultural
- Barriers exist, infrastructure can help
- Interoperability

    - Wizard systems guide users based on selected semantic terms
    - Run and deposit analytical workflows

- Reproducibility
    - Data is well-described, open, and DOIs will be minted
        - Data citation
    - Programmatic access to all layers
- How can we help researchers realise the benefits of sharing data?

Discussion Group 1+2 - Assess and find ways to increase participation from universities, government and research organizations in the agricultural sector worldwide

- Hugo Besmer Information specialist - plant protection
- Mark Goovaerts (Librarian, repositories)
- Kana Vitner (FAO since 2012)
- Imma Subirat
- Patrice Ajai-Ajagbe (Commonwealth)
- What's in it for students and scientists
- godan.info
- Research data management group for Godan
- Global open data for nutrition
- Institutional issues
- Deliverables over the next 6 months

- Set up a working group for training activities
    - Spanish course can be used by us, the users
    - User's cases can make it more agricultural in flavor
    - "IMARC" platform for courses, very much used
- RDA no legal entity, catalyst
- Look for funding elsewhere
- RDA is for scientists
- It is a forum for exchange
- Working group never stand alone, always under an interest group
- Private sector not so much in RDA; GODAN is interested in private/public relationships
- ISPA, ASA, CSSA, SSSA should reach out to RDA
- USDA, AAFC
- USDA strong in GODAN
- NSF is behind RDA in the USA

Report by Group 3

- Wheat can be expanded to "cereals" including corn
- Recommendation for repository
    - Public existing repository: free access and free processing
- Dataset publication with DOI
- Data types
    - Genetical variance
    - Genomic variant
    - Genomic data
- Visualization of large datasets (softwares available)
- Germplasm
    - DOI are the best option for identification, but come with a small cost
- Phenomics
    - Environmental ontologies
- Metadata aspect

Report from Group 4 - Interoperability (general)

- Generic
    - Name authority
    - Catalogues of datasets
    - Registries
- Data and metadata quality to the correct use of data standard
    - Validation mechanism
- Disambiguation
    - Traits and organisms are more specific to agriculture
- Standardization variables, units of measures
- RDA role is bring everyone together. Inform on what is going on. New services, new standards, etc.

Report from Group 5 - Soil

- Who needs the soil data? Many groups.
- Ownership of soil data: national and international levels
  - Global soil partnership
- Problem about distributing data on private property (soil)
- Communication, dissemination
- Metadata structure
- Quality control/quality insurance

## WORKSHOP ON BIG DATA FOR FOOD, AGRICULTURE AND FORESTRY: OPPORTUNITIES AND CHALLENGES

***http://www.big-data-europe.eu***

**INRA's Big Data perspectives and implementation challenges (Pascal Neveu, IMR MISTEA, INRA, Montpellier)**

- Data challenges in science
  - More data production
  - Experimental datasets available on the Web
  - More collaborative and integrative approaches
  - Management, sharing and data analysis play an increasing role in research
  - Discover, combine and analyse these data
  - Big Data: when traditional methods stop to work
- Volume, **variety** **(complexity)**, velocity, **validity**, veracity
  - Specific for agriculture
- Production of a lot of heterogeneous data for understanding
  - Open new insights
  - Allows to know which theory are working in practice
- Phenotyping
  - 2013: 40 Tb
  - 2014: 100 Tb
- Extremely diverse data
  - Web API, ontology sets, NoSQL and semantic web methods
  - Need data cleaning (automatic diagnose and management)
- High throughput phenotyping
  - Hard to produce
  - Hard to manage
  - Also hard to analyse

**Big data opportunities for marketing of horticultural products (Tim Verhaart, Wageningen)**

- Weather is important because it impacts supply for the market

- Consumption, production, trade, news media
- Data are present in public administrations, but not open
    - Gvt promote Open Data but at a lower level many obstacles must be overcome
    - Concerns about competitive relations
- Semantic heterogeneity (in product classifications, etc.)
- Many SMEs in this sector have limited capacity to invest in the IT required
- BIGt&u project
    - Infrastructure for market information
- Work packages
    - Semantic technologies
    - Social media filtering
    - Metadata
    - IT architecture and development
    - Social, legal, financial aspects
    - Application prototyping and community building across horticulture and IT sectors
- Goal: efficient, uniform access to a wide variety of sources for market data

**Big Data in CGIAR (Élizabeth Arnaud, Bioversity, France)**

- Also in developing countries
- Big Data for climate smart agriculture
    - CIAT analyses large, real-world data sets from annual survey on rice to produce recommendations much more quickly
    - Result: Identified the most productive rice varieties and planting times for specific sites and seasonal forecasts. Recommendations could potentially boost yields by 1 to 3 tons per hectare.
- Big Data: look for patterns
- Crowdsourcing varieties
    - 500 farmers per site will be given 3 blind varieties in small quantities to be tested under their own conditions
    - ClimMob
- CGIAR Big Data
    - Large amounts of data published
- 8 agrifood system research programme
- 5 integrative programmes
- Call for projects
- IFPRI-led EOI: tools for driving interdisciplinary and collaborative Big data analysis
    - Cases:
        - Scalable satellite-based crop yield mapper
        - Crop water
- Originates from US

Nikos Manouselis (Agro-Know) on food safety data use cases

Interactive Session (Identification of relevant data sources, projects/initiatives, etc;

- Remote sensing, proximal

Identification of main challenges and problems

- Availability
  - Private
  - Need historical data
- Need data scientists

Note: Copernicus satellites produce 4Tb of images free, every day

**Big Data challenges and solutions in agricultural and environmental research (Rob Lokers, Alterra, Wageningen)**

- Historic perspective
  - Crop science
  - 1960-2080 = Institutional data collection
  - 1980-2000 = First computer models
  - 2000-2010 = Institutional applications
    - Integrated modelling frameworks
  - 2010-2015 = IT improvements (meta data, semantics)
    - Open data across sectors
  - 2015-2020 = Big data: one massive linked data pool across disciplines and strong computational capabilities
    - Computional capabilities
    - New data sources
      - RS, crowd sourcing, rapid phenotyping/omics, social media
    - Potential to solve problems!
- Expectations for the (near) future
  - Pyramid
    - (Big) Data
    - Information
    - Knowledge
    - Wisdom (policy makers, industry, societal stakeholders)
- Some examples
  - MARS (Monitoring Agricultural ReSources)
    - Weather archives, live data strems, soil/crops DB, models
  - Smart farming: monitoring, planning & control
    - Crop/animal level, farm level, market level, environmental level
    - Hard to combine data in such context
- Technologies and RS
  - Technologies used
    - RDBMS, geo-databases
    - Various "old & proven" programming languages (esp. for modelling, data processing)
    - RS: dedicated tools (ENVI, R, GDAL, etc)
    - Harmonized info/data models, but still per discipline
    - High performance clusters/grids
  - Experimental (ICT research)
    - RDF DB
    - Vocabularies and ontologies (no alignments)
    - NLF algorithms

- Expectations vs reality in 2015
    - New tech solutions (RDF db, ontology alignment, NLP)
        - Successful initiatives using hybrid solution, often build on "proven" technologies
    - "Magical" semantic (and linguistic) query processing
        - "Technical" query processing (e.g. through SPARQL) - gap still quite big
    - Transparent to big, distributed, heterogenous datasets
        - Many success on metadata level and biblio, cumbersome first attempts to harmonize heterogenous data streams
        - Custom-build data colletion and processing chains still remain dominant

**A global linked and open data infrastructure for agricultural development (Valeria Pesce)**

- Infrastructure on Big Data, new technology platform
- Big Data Europe proposed infrastructure
    - ICT + computing
    - Re-deployable generic infrastructure with "n" domain-specific big data integrator instances
- Ag-datacontext
    - Actors have built several infrastructural components over the years
    - Mostly vocabularies, ...
    - Little work on computational services
- Positioning
    - Interlink between not-so-big and big data
- Registry for datasets and vocabularies
- Would like the new big data infrastructure to be interlinked with existing infrastructural components
- Will have to link with "Elixir" (life science information)
    - https://www.elixir-europe.org/
- Would like to federate registries as "one stop" portal

**(Stephan ...? représentative of the European Commission)**

- Researchers do not share the data?
    - In Horizon 2020: data management is a requirement
- Avoid duplication of initiatives and tools
- Horizon 2020 draft work progammes 2016-17 just released
    - Where is the next area in Europe where data will be central to activities? Could it be agriculture? Reingineering of an entire production sector around data management.

**Big Data Europe (Soren Auer, Big Data Europe Coordinator)**

- The "V" for "variety" is often neglected
- Windshield wiper as rain sensors for micro weather prognosis. Automotive industry can become data provider for agriculture and other industries...

- Prescriptive analytics suggest decisions
- Agriculture does not have much different requirements than other sectors
- Zoo of different open source big data applications
    - Want to help sorting out and help selecting
- Analysis of historical data
- Analysis of actual data with low latency in "real-time"
- Interactive analysis by online users
- Data aggregator platform for both semantic and non-semantic data
- W3C Interest group: "Food & Agriculture"
- Webinars, community group

    - W3C - semantic web (Link you data)

## RDA 6TH PLENARY

**Opening ceremony**

- 600 registrants
- 34 countries

Mark Parsons RDA Secretary General & Patrick Cocquet

European Community representative

- Data management plan part of funding agreements with beneficiaries
- Translate work into deliverables: urgent to match Horizon 2020 data production expectations

European Commissionner Gunther Ottinger (video message)

- European cloud initiative, including science cloud
- Copernicus satellites, temporal & spatial resolution + data continuity, freely available
- Massive data must be exchanged, + security, + protection

Mark Parsons

- Climate change issue, big data
- Industry has to be part of the solution. Solution with people, not for people
- Social + technical bridges
- Trust is central in data sharing
    - Comes from shared experiences and perspectives
    - Data reuse and adoption

- RDA: Neutral forum where we can address issues
- Work from use cases, need a diversity

Axelle Lemaire (Ministry of State, Digital Sector, France)

- Data circulation, used and reused
- Only then: increase in value
- France, 3rd in the world for data openess
- www.data.gouv.fr
- http://www.data.gouv.fr/fr/topics/agriculture-et-alimentation/
- International standards needed

Unleashing the power of earth observations - Together / Barbara Ryan (GEO Group on Earth Observation)

- Importance of preserving archives to learn from the past
- 87 participating organizations
- Metadata very important
    - Compatible to other systems
    - Brokering approach
- GEO
    - Brokering organization; does not own data
- LANDSAT
    - Before open-data policy: 53 scenes/day
        - Bought by government or agencies... not a good investment for the government
    - After open-data policy: 5,700 scenes/day
        - China jumped in quickly! Nice products now.
        - Annual economic benefits
            - USA 1.70 B$
            - International 400 M$
- RADARSAT (Canada)
    - Same effect of open data
- Uncertainty is the enemy of investment; shared investment is the antidote to uncertainty
- Crop Information for Deciion-Making
    - GEOGLAM part of G20 actionplan on food price volatility
    - New crop outlook
    - Rice crop monitoring
    - GEOGLAM crop monitor for AMIS, operational since 2013
- Countries have borders, Earth observations don't
- GEO Week, Mexico city in November 2015
- www.earthobservations.org

**Working Meeting Sessions**

**BoF on Convergences in Archives, Records management and Research Data Curation**

- http://bit.ly/1KzfvTf

**Journal Research Data Policies**

- 400 journals policies, about 22% only had strong committment for datasets publications
- Jisc Journal Research Data Policies Registry
- What data? Only data behind the paper? Additional data as well?
- Data sharing is not data deposit
- Publisher level vs Journal level; need to align
- Creating standards for policies would be more important than creating a repository of policies
- Impose something on researchers
- Plos makes data publication compulsory
- Funders should ask for the initial data set
- Journals should ask for the data set necessary to reproduce the content of the article

**Joint meeting of IG Data Rescue, IG Geospatial, IG Big Data Analytics, IG Domain Repositories & IG Libraries for Research Data**

- Access all data that can be accessed
- Long stamp: benefit
- Digital: only 30 years at best; but 120 years available!
- Cases studies
- Proposal for online guidelines about data rescue: how to. Synopsis

**CERN in Switzerland**

- jamie.shiers@cern.ch
- All data is at risk
- CERN accelerator complex: data is available publicly
- All worldwide collaborations involving several hundred people per "experiment"
- LEP "data" rescue 2015
    - Was big data at the time (1989 - 2000)
    - 500TB now 2 tape copies + 1 disk copy at CERN
    - Copies exist also outside institutes
    - Physics output (papers, presentations, PhDs) continues
    - Preservation = Data + Documentation + Software
        - + Knowledge but that is harder to define
    - We measure the cost and the benefit(s) - mandatory for funding
- Data should be useable 30 years after final data taking; data itself should be carried forward
- But constant change "under your feet" requires continuous validation an possibly migration...
- Some degree of luck for CERN to preserve its data
- Long term data preservation cannot be built on luck

- There are probably disruptive changes that will be hard to overcome with specialist knowledge

**Sea level data archaeology (National Oceanography Centre)**

- Analogue records sometimes available
- 1871 - 1926 in Malta: photographed tidal charts
    - Tsunami recorded in 1906
- Crowdsourcing? www.oldweather.org
- Metadata and discoverability
- Main barrier: lack of funding
- Scientists must actively market data rescue and digitalization

**Data rescue award in the geosciences (Kerstin Lehnert)**

- Rescuing Legacy Data For Future Science
- GeoResJ
    - 22 papers in 4 categories of rescue
- 5 k$ price

**Agriculture Data Interest Group (IGAD): Results of the IGAD premeeting RDA P6 : Moving forward**

- Register to the wiki
- Goals
    - Promotion of good practices
    - Platform for networking
    - Promotion of interactions and projects among the major international institutions and groups worldwide
- Increase participation from univ., Gvt and research organizations
- Group 1 Hugo Besemer
    - Private sector has to be in the loop
    - GODAN, possible cooperation
        - http://www.godan.info
        - http://www.godan.info/about/
- Group 2 Cyril Pommier and Pierre Larmande
    - Increase data access and availability
    - http:/wheatis.org/
    - Metadata
    - Ontologies
    - Tools
    - Formats
    - Use case
    - Data types
- Group 3 John Fereira (Cornell)
    - From infrastructure to service
    - Mapping of vocabularies

    - Name authorities
    - Disambiguation
    - Standardization of variables and units of measure
    - Catalogs of datasets
    - Automatic annotation
    - Disambiguation
    - APIs
    - Integrating heterogeneous data
    - Linking between different types of data
    - Decision-making ontologies
    - Process/workflow modelling
    - User interfaces of annotating
    - Interoperable data / dataset management tools
    - Support for publishing datasets
    - Metadata
    - Registries
    - Data/metadata quality
    - Data ownership
    - Traceability
    - Business rules
    - Reasoning
    - Use rights
    - Provenance
    - Narrowing down
        - Data / metadata quality
        - Disambiguation
        - Standardization of variables and units of measure
    - Other things are already taken care of
- Group 5 Pandi Zdruli and Devika Madalli
    - Establish an "Interest group on soil" inside the RDA