Making Open Data Work for Research and Researchers

Sabina Leonelli
Exeter Centre for the Study of Life Sciences (Egenis)
& Department of Sociology, Philosophy and Anthropology
University of Exeter
@sabinaleonelli

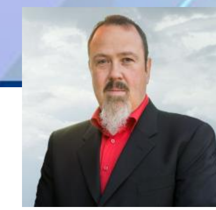
MLE on Open Science

Jan 2017 – Jan 2018, 13 countries

Frank Miedema
Chair

Katja Mayer
Rapporteur

Sabina Leonelli
Expert



Kim Holmberg
Expert

Three topics:

1. The potential of altmetrics to foster Open Science
2. Incentives and rewards for researchers to engage in Open Science activities
3. Guidelines for developing and implementing national policies for Open Science

https://rio.jrc.ec.europa.eu/en/policy-support-facility/mle-open-science-altmetrics-and-rewards

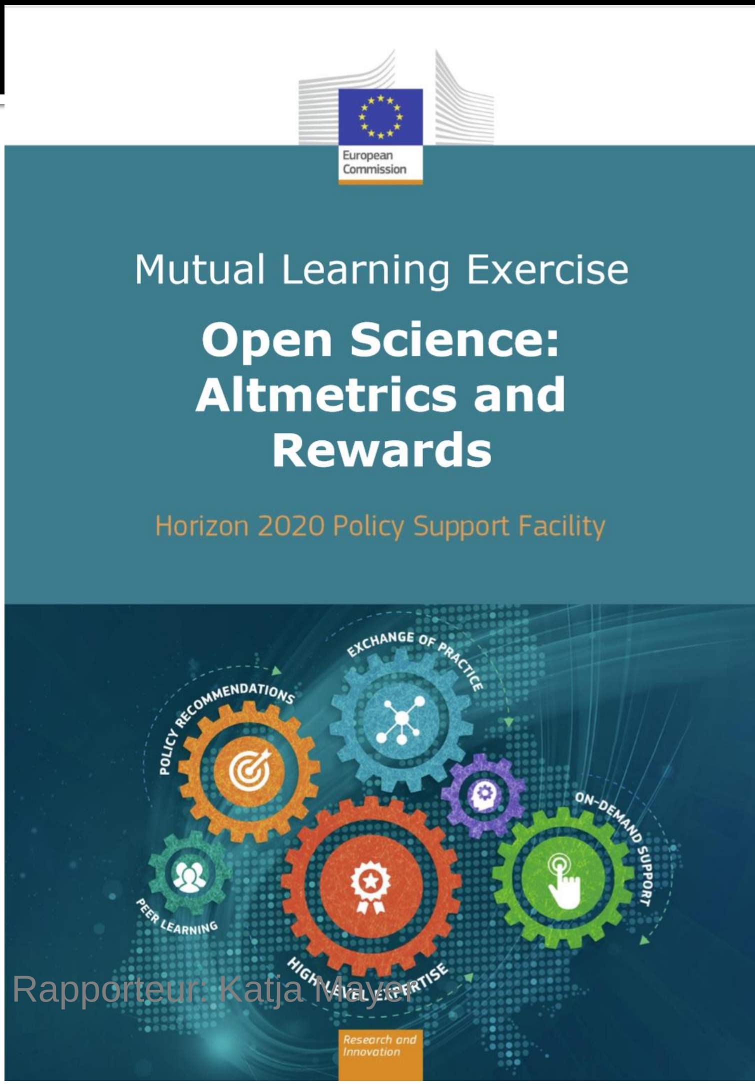
Armenia
Austria
Belgium
Bulgaria
Croatia
France
Latvia
Lithuania
Moldova
Portugal
Slovenia
Sweden
Switzerland

MLE on Open Science: 4 Thematic reports

Different types of Altmetrics (Kim Holmberg)	**Altmetrics and Rewards** (Kim Holmberg)	**Incentives and Rewards to engage in Open Science Activities** (Sabina Leonelli)	**Implementing Open Science: Strategies, Experiences and Models** (Sabina Leonelli)
Conclusions: • Altmetrics are not yet being used for research evaluation purposes. • Altmetrics hold a lot of promise, but it is too early to use them for research evaluation and decision making. • More research is needed.	Issues are: • Not enough evidence • Limitations of (proprietary) data sources • Methods are not yet open	The report suggests that incentives and rewards should be applied to three groups of key stakeholders: (1) researchers; (2) research-performing institutions and funding bodies; and (3) national governments.	This report • proposes a National Roadmap for the Implementation of Open Science • outlines key priorities and principles underpinning the implementation of Open Science at the national level • reviews existing experiences in developing and supporting OS activities and related policies • summarises the strategies, lessons learnt and models

http://europa.eu/!bj48Xg

Final Report

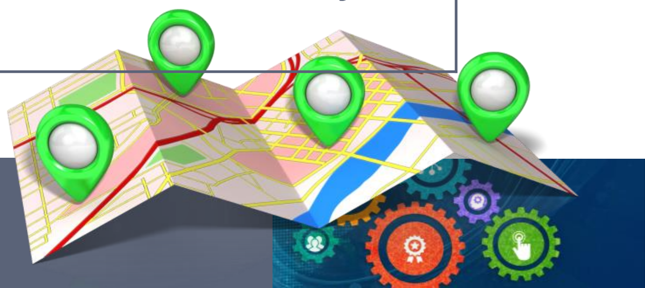


Rapporteur: Katja Mayer

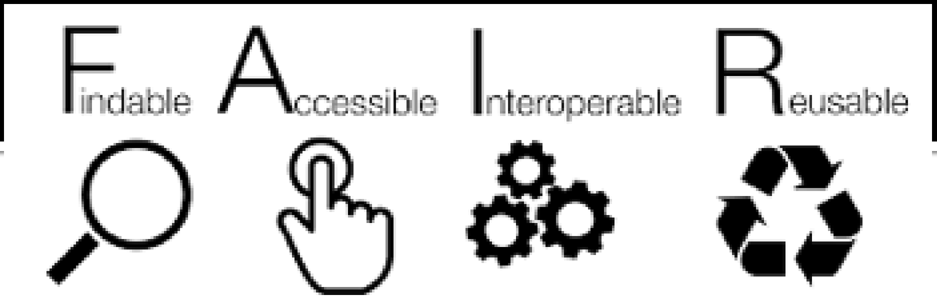
1. INTRODUCTION
2. METHODOLOGY
3. BACKGROUND OPEN SCIENCE
   - The status of Open Science in Europe – implementation and aspiration
   - Altmetrics
   - Incentives and rewards
   - National initiatives for open science
4. POSITIONS AND PERSPECTIVES FROM MEMBER STATES AND PARTICIPATING COUNTRIES
5. LESSONS LEARNED
   - Key concerns and best practice
   - Priorities
   - Roadmap for the implementation of Open Science
   - Conclusions and Next Steps

Roadmap for Open Science Implementation

Map	Identify key stakeholders and Open Science champions
Plan	Devise national strategy through consultation with stakeholders
Incentivize	Change reward system to incentivize all aspects of Open Science
Promote	Encourage critical and informed thinking
Support	Participate in international initiatives
Implement	Implement strategy, starting from Open Access
Monitor	Monitor and tackle emerging issues as they arise



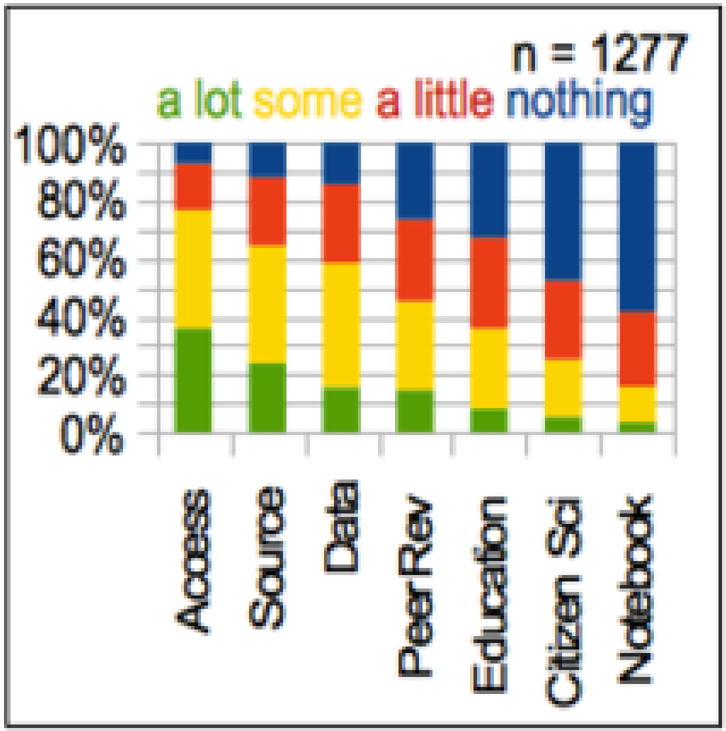
The key role of data curators and infrastructures

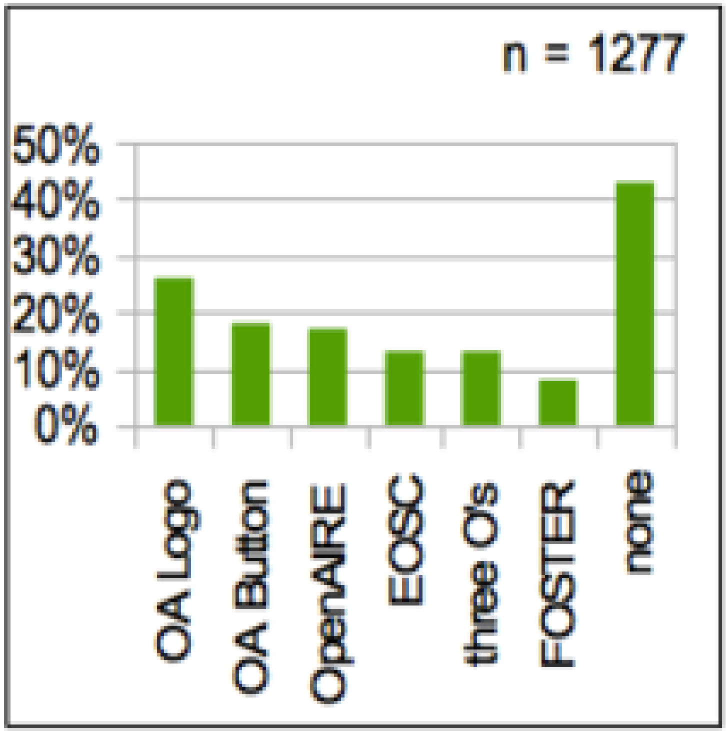


- Making data FAIR requires
  - coordination and interoperability of data infrastructures
  - making data *mobile* and *useful as evidence* across sites, contexts, uses
  - making data infrastructures trustworthy and user-oriented
  - ensuring the fairness of data handling and implications

- **Major challenges to realising that potential**

Low Awareness of OS Activities and Tools





Source: EU Working Group on Education and Skills under Open Science, 2017

Complexity of tools and skills required to make data FAIR

Data management and best practice for plant science

Sabina Leonelli, Robert P. Davey, Elizabeth Arnaud, Geraint Parry and Ruth Bastow

Table 1 | General tools for data management.

Type of tool	Function	Examples of relev
Open lab books	Digital and shareable version of traditional lab books	RSpace (http://wv
Generic open data repositories	General storage for many different data types	Figshare (http://w DataVerse (http://www.dataverse.org)
Specific databases	Fine-grained datasets that require subject-specific metadata	The Arabidopsis Information Resource (http://www.arabidopsis.org) The Bio-Analytic Resource for Plant Biology (http://www.bar.utoronto.ca) iHub (http://www.ionomicshub.org/home/PiiMS)
Data portals	Aggregating and providing visibility for various databases and resources	Araport (http://www.araport.org) Biosharing (http://www.biosharing.org) Agroportal[22]
Bio-ontologies	Keywords for the annotation, ordering and retrieval of data	Plant Ontology[15] Crop Ontology[21]
Metadata standards	Standardization of experimental data collection	Minimal Information on Biological and Biomedical Investigations (http://www.biosharing.org/standards) Minimal Information about a Microarray Experiment[27] Minimal Information about Plant Phenotyping Experiments (http://www.cropnet.pl/phenotypes/?page_id=15)
Identifiers for research materials	Annotation and retrieval of research materials on which experiments were originally performed	Germplasm Resource Information Network – Global (http://www.grin-global.org/) Multi-Crop Passport Descriptors (http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v2-mcpd-v2) Genesys (http://www.genesys-pgr.org)
Informatics standards	Software tools helping to format, store and visualize data	Breeding API (http://www.docs.brapi.apiary.io/) InterMINE (http://www.intermine.org)
Data annotation pipelines	Annotation of data from generation to reuse	Integrated Breeding Platform (http://www.integratedbreeding.net/) CropStore (http://www.cropstoredb.org/description.php) eDal (http://www.edal.ipk-gatersleben.de)
Guidelines of good practice	Articulation of data management principles and actions fostering data reuse	FAIR Data (http://www.force11.org/group/fairgroup/fairprinciples) Wheat Data Interoperability Guidelines[31]

(source: Leonelli et al 2017, *Nature Plants*)

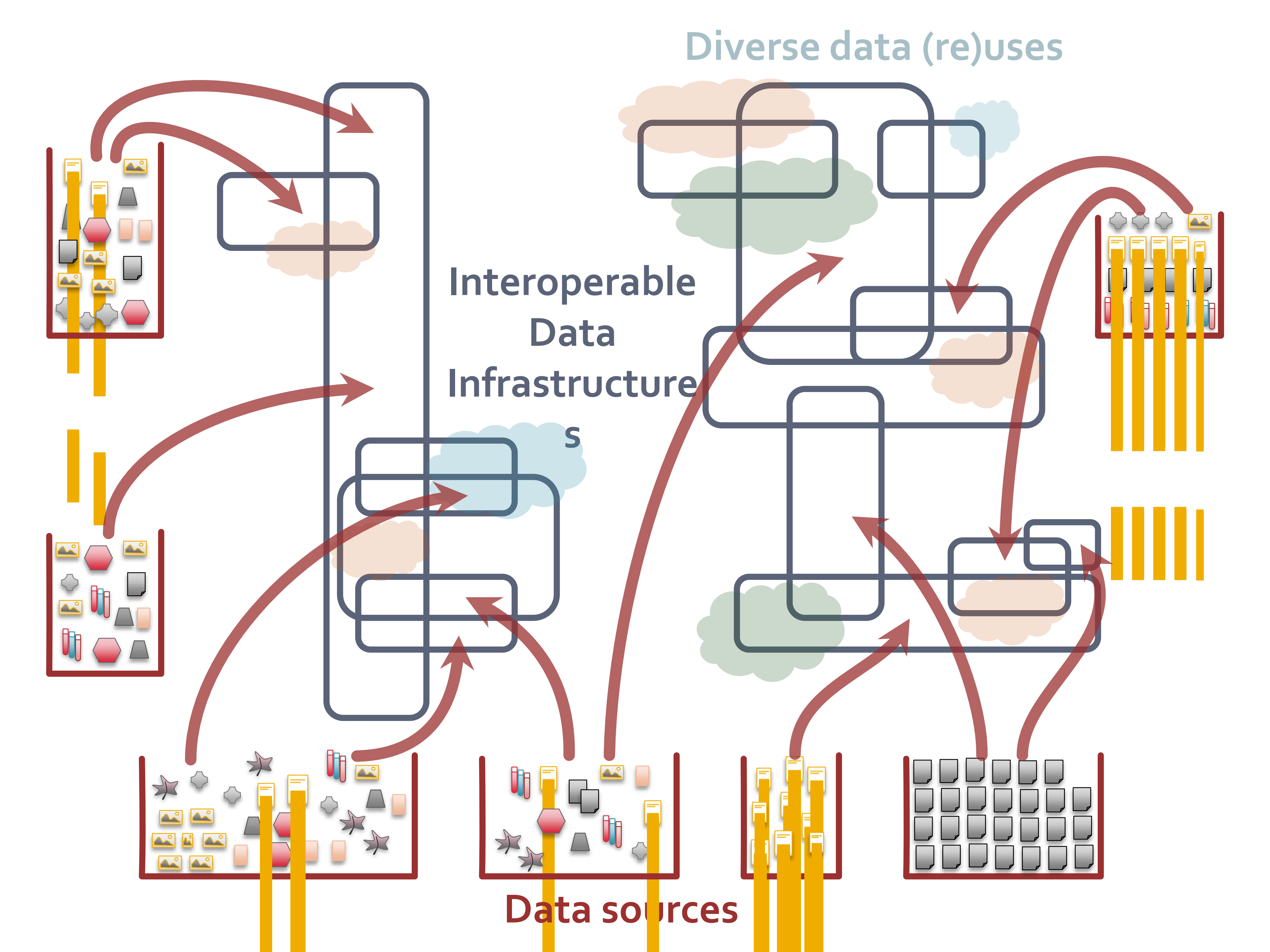
Confusion Among Researchers Over

- **What openness means in practice**
  - Some common interpretations: “free of license”, “free of ownership”, “under CC-BY license”, “common good”, “good enough to share”, “unrestricted access or use”, “accessible without payment” (Grubb & Easterbrook 2011; Levin, Leonelli et al 2016)
- **How can it be implemented**
- **What is legal** (how does openness apply to commercially or security sensitive research?)
- **What is ethical** (how to protect individuals & groups from harm?)
- **What is recommended by whom** (funders, learned societies, publishers, research institutions, governments..)

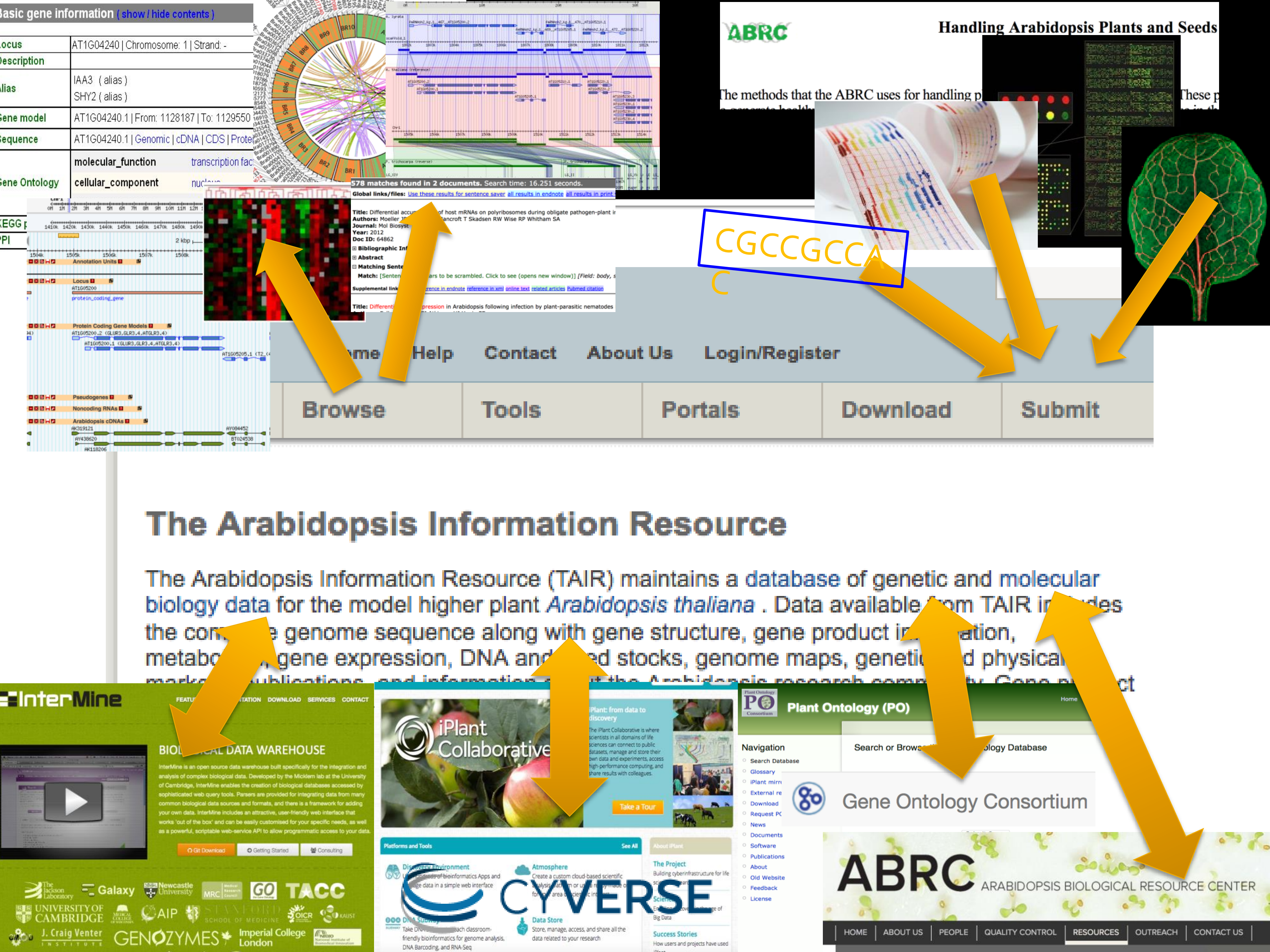
DATA_SCIENCE project:
The Epistemology of Data-Intensive Science

Tracking **data journeys**
To understand how data move from sites of *production* to sites of *dissemination* and *interpretation/use,* and with which consequences

- **Approach:** philosophy, history and social studies of science
- **Focus:**
  1. **Databases** as windows on material/conceptual/institutional labor required to make data widely accessible and useable
     - labels & software to classify, model, visualize, retrieve data
     - management of infrastructure and communications
  2. **Data re-use cases** to investigate
     - conditions under which data can be interpreted
     - implications for discovery & what counts as good research
     - role of Open Science movement in knowledge generation


Diverse data (re)uses
Interoperable Data Infrastructures
Data sources

Green research – plant science and food security



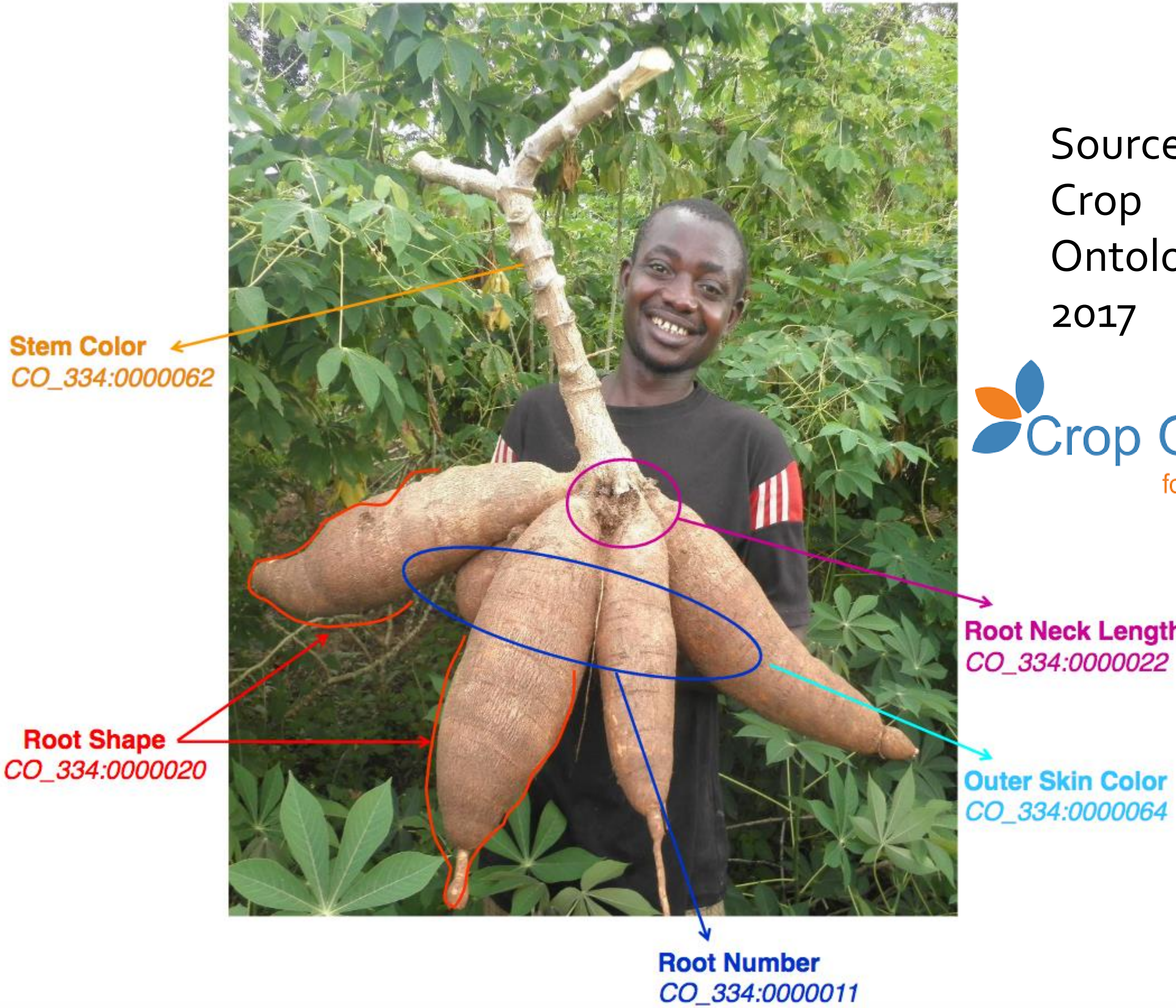
Basic gene information (show / hide contents)
Locus
AT1G04240 | Chromosome: 1 | Strand: -
Description
Alias
IAA3 (alias)
SHY2 (alias)
Gene model
AT1G04240.1 | From: 1128187 | To: 1129550
Sequence
AT1G04240.1 | Genomic | cDNA | CDS
Gene Ontology
molecular_function
cellular_component
578 matches found in 2 documents. Search time: 16.251 seconds.
ABRC
Handling Arabidopsis Plants and Seeds
CGCCGCCAC
Home
Help
Contact
About Us
Login/Register
Browse
Tools
Portals
Download
Submit
The Arabidopsis Information Resource
The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana . Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.
InterMine
BIOLOGICAL DATA WAREHOUSE
iPlant Collaborative
CYVERSE
Plant Ontology (PO)
Navigation
Gene Ontology Consortium
ABRC
ARABIDOPSIS BIOLOGICAL RESOURCE CENTER
HOME
ABOUT US
PEOPLE
QUALITY CONTROL
RESOURCES
OUTREACH
CONTACT US

High Throughput Phenotyping

Six phases involved in data journey from production to analysis:

[1] Preparing specimens
[2] Preparing and performing imaging
[3] Data storage and dissemination
[4] Coding for analysis
[5] Image filtering
[6] Image analysis
[6] Calibration and further analysis

National Plant PHENOMICS Centre



Source: Crop Ontology 2017

Crop Ontology
for agricultural data

Digitalisation of Phenotype Workflow

cassavabase

Field Layout & trait file

Fieldbook (tablets, mobiles)

Curated data

Data export

Data collection

Network folders on IITA server

Red research – human health and wellbeing

Medical insights from non-human research: cell regulation in yeast

PomBase

The scientific resource for fission yeast

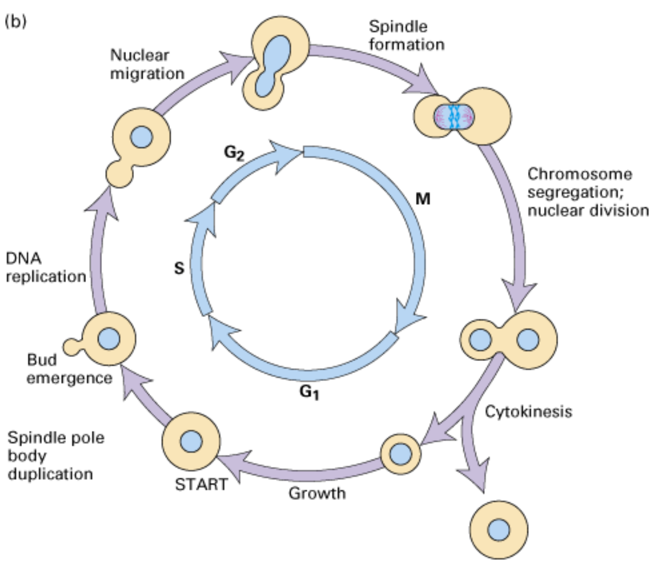
Search ▾ Submit

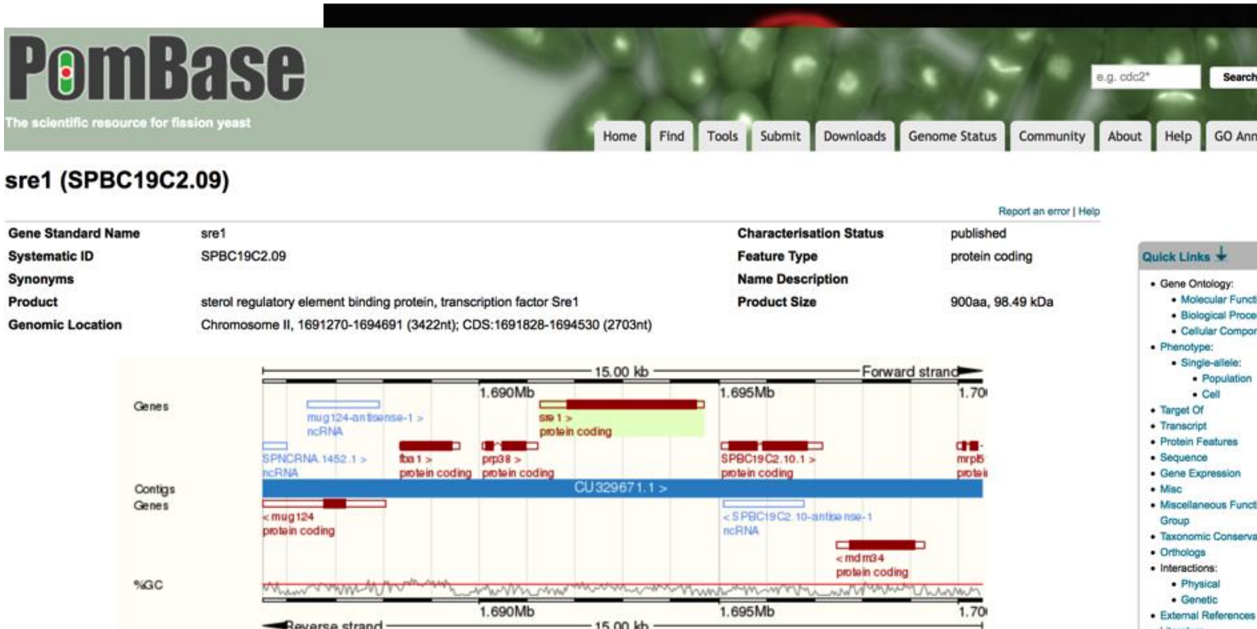
Welcome to PomBase

PomBase is a comprehensive database
Schizosaccharomyces pombe, providin
functional annotation, literature curation
scale data sets

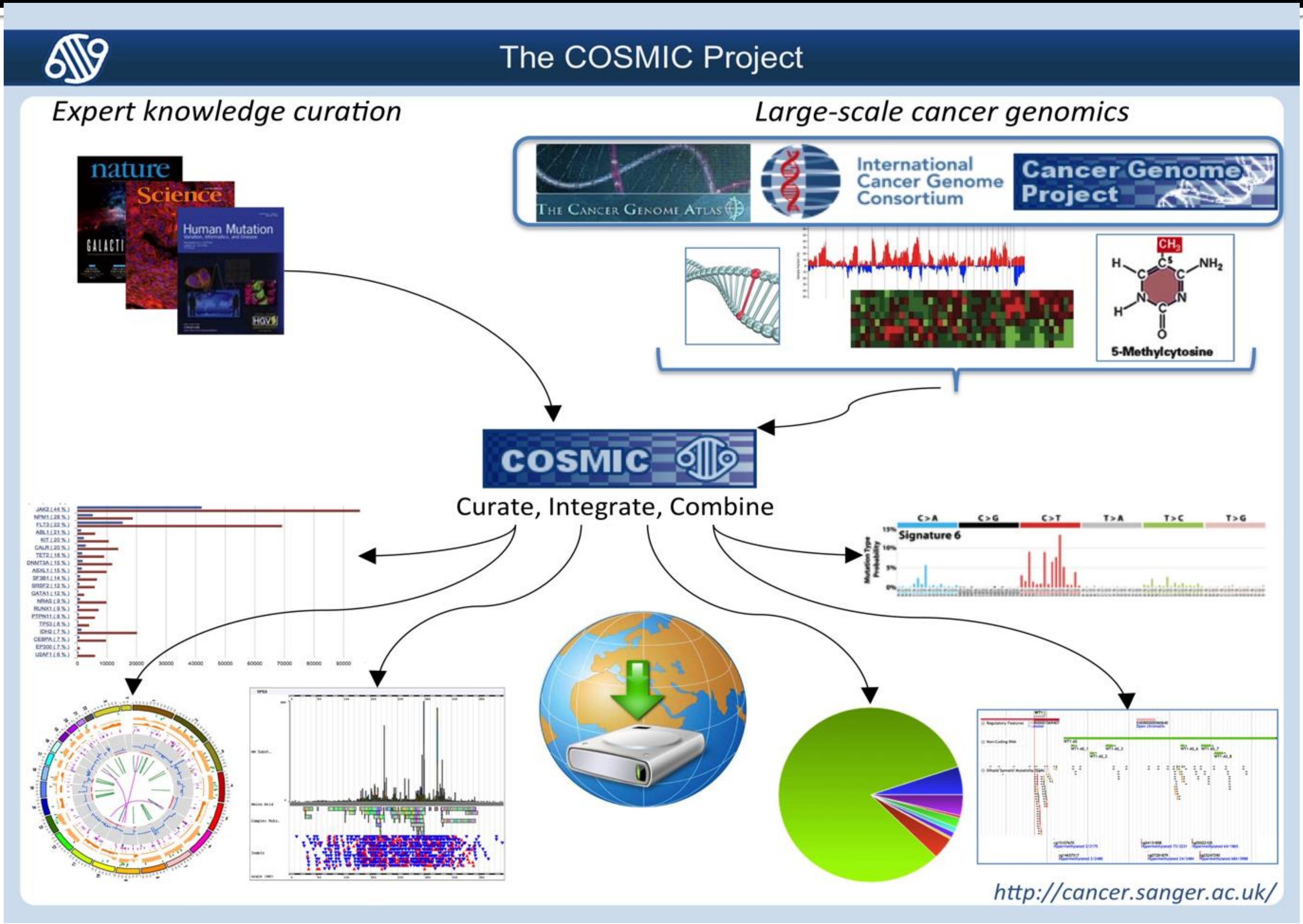
SGD *Saccharomyces* GENOME DATABASE

Analyze

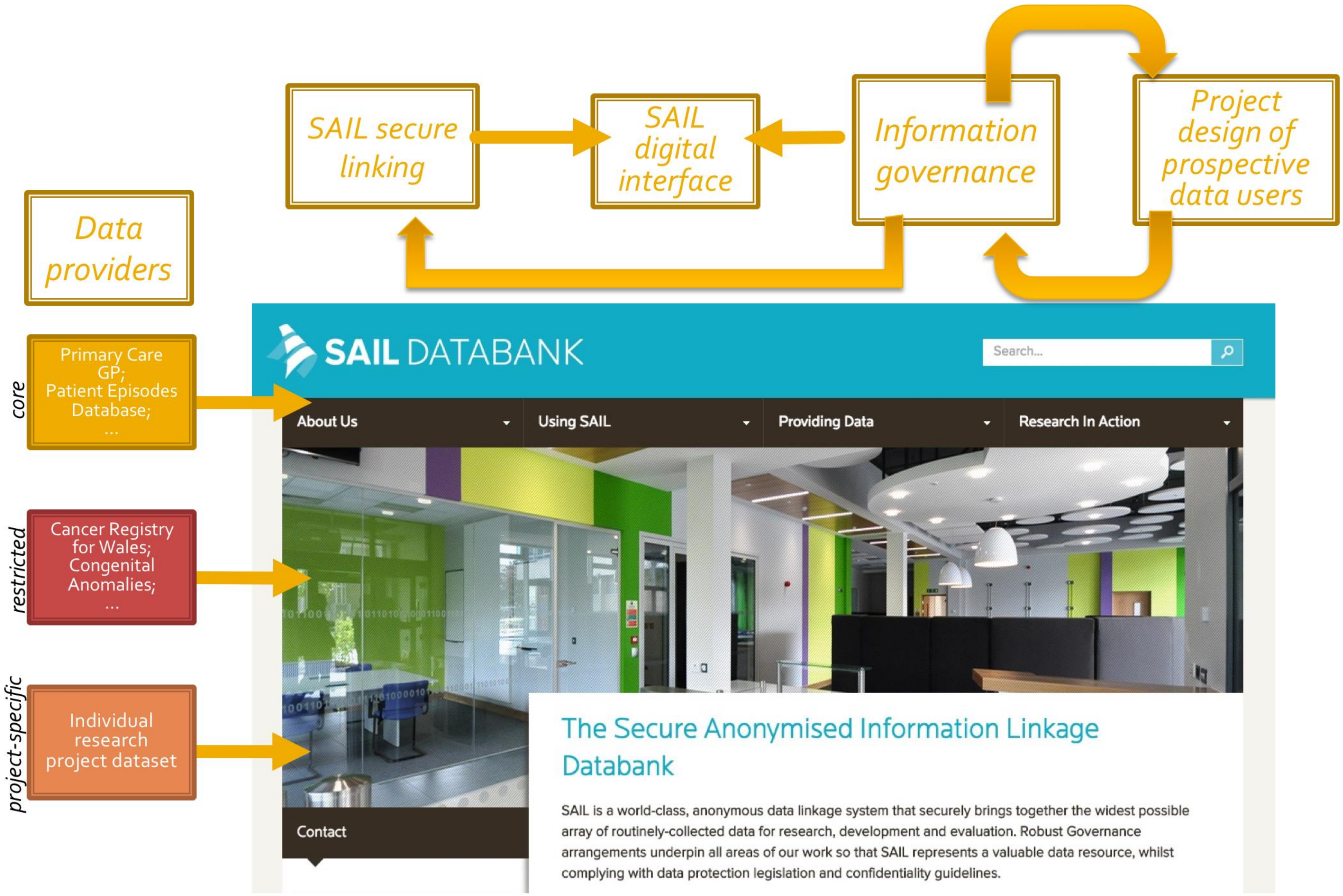




From somatic mutations to clinical assessment

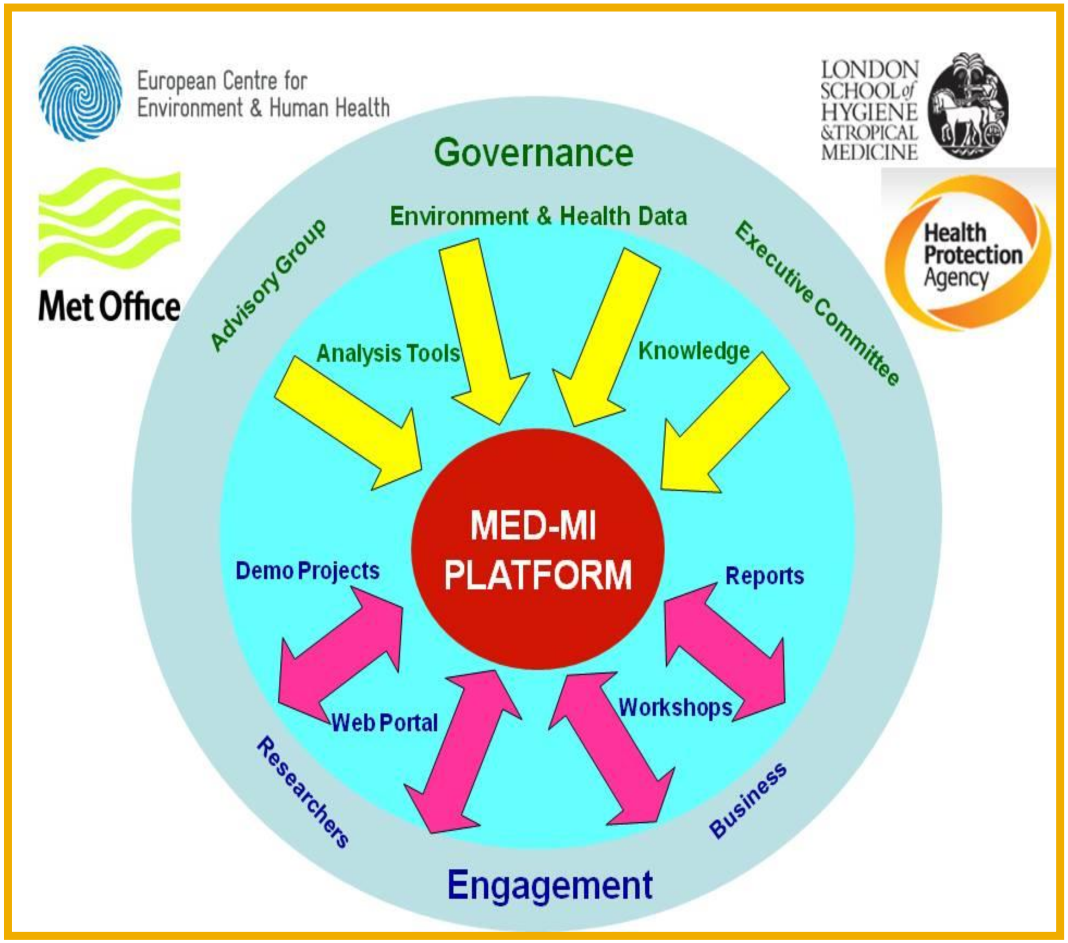


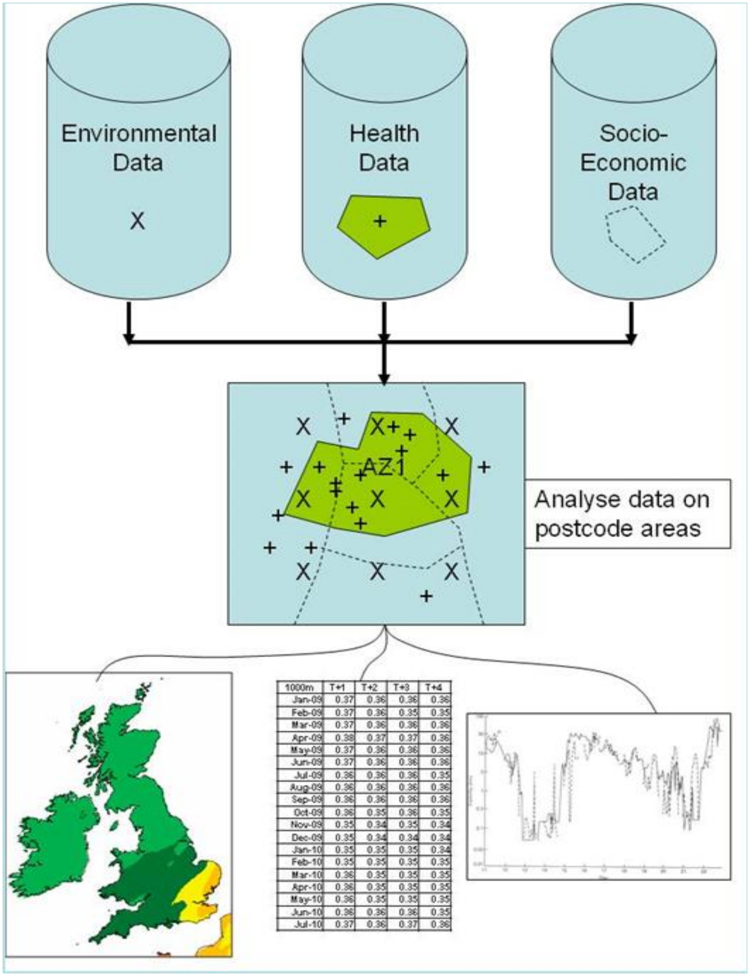
From clinical assessment to data re-use for research



Bringing green and red together: human and environmental health

Medical and Environmental Data Mash-up Infrastructure





Lessons learnt

Epistemic Trouble with Online Data Sources

- Research data collections available online represent highly selected data types from a small proportion of available sources
- Selection based on *convenience*, *tractability* of the data and political-economic *conditions of data sharing*, rather than on epistemic choices
- Peer reviews structures unclear and often lacking
- Misalignment between IT solutions and research questions/needs/situations (e.g. problems with access to software)
- No sustainable plans for maintenance and updates of most data infrastructures
- No sustainable plans for tracking and accessing related samples/materials

Lesson Learnt 1: *Context-Specific* Curation Is Key to Data Reuse

- **Field-specific data curation is essential** to data re-use and interpretation, yet badly underestimated and not rewarded
  - Do not throw the baby out with the bathwater: value of long-standing research traditions and reviewing methods
  - Crucial to remain user-friendly and fulfil expectations of users
  - Need case-by-case judgments on research quality and fruitful modes of data sharing
  - Pluralism in methods and standards contributes robustness to data analysis, and reduces risk of losing system-specific knowledge

Lesson Learnt 2: Long-Term Maintenance Is Key to Trustworthiness

- Regular updates across nested infrastructures
- Business plan for long-term sustainability
- For OS, this means:
  - Clear relation between international field-specific databases, international clouds, national clouds, institutional repositories
  - Make sure each node is resilient and system is not crippled by individual node failure (now all independently funded, typically in the short term)

Lesson Learnt 2: Long-Term Maintenance Is Key to Trustworthiness

- Particularly important since hard to guarantee data quality
- Criteria for what counts as good data – or even as data altogether – vary dramatically even within the same field
- Role of confidence assessments on data quality and reliability (again: field-specific curation is key)

Lessons Learnt 3: Which Data and Why?

Substantive disagreements over data management:

- Methods, terminologies, standards involved in data production and interpretation
- What counts as data in the first place (data as a *relational category*, Leonelli 2016, 2018)
- What counts as meta-data

Lessons Learnt 3: Which Data and Why?

- Re-use often linked to participation in *developing* data infrastructures

→ rarely the case for busy practitioners, considering also gap in skills

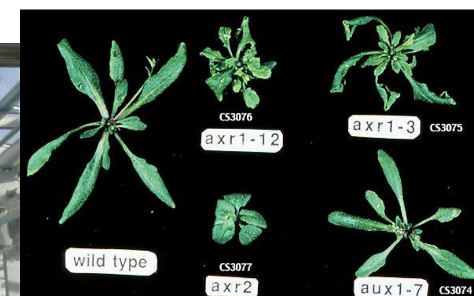
- Indiscriminate calls for open data can lead to serendipity in what data are circulated and when

→ Need explicit rationale around priority given to specific data types and sources (e.g. 'omics' in biology)

Lessons Learnt 4: Data and Materials

- Sharing of related materials via reliable stock centers and collections: *rarely* available & coordinated with databases

- E.g. model organism stock centres, biobanks




DiSSCo
Distributed System of Scientific Collections
Home
News
Partners
Projects
Network
Files
About Us
114 European Museums
21 Countries
Largest ever agreement to unify European natural science collections

Lessons Learnt 5: Role of Ethics, Humanities & Social Sciences in Data Management

- **Ethical, social and security concerns** increase quality and re-usability of data/infrastructures
  - Related skills are as central to data science as computational skills
  - Data re-use requires well-informed, sustainable, inclusive, participative development of data infrastructures

- **Open Data and Data Science training**: Data science is not a branch of engineering, but rather requires input from all fields, esp. social science and humanities

Conclusion: No reliable and effective (Open) science without trustworthy data curation

- Effective, context-specific curation
- Sustainability and maintenance
- Built-in ethical safeguards, social relevance and resilience
- Robustness (plan B if specific standards/services fail)
- Criteria for data and meta-data inclusion and formatting
- Clear link to samples and specimen collections

Conclusion: building on the French National Plan for Open Science

- Clear national commitment and institutionalization (Chief Data Officer)
- Promises to ease legal complexities of Open Data
- Promotes shift in research evaluation to recognize Open Science: data curation deserves specific attention!
- Promotes Open Data while recognizing disciplinary diversity
  - Welcome emphasis on role models and OS champions

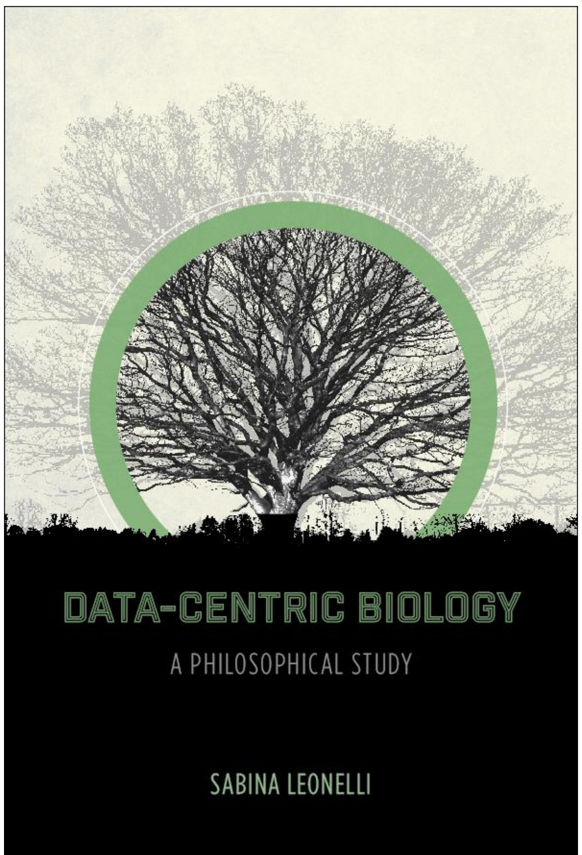
Areas of immediate concern to French plan

- **Research data collections available online represent highly selected data types from a small proportion of available sources**
- Selection based on *convenience*, *tractability* of the data and political-economic *conditions of data sharing*, rather than on epistemic choices
- Peer reviews structures unclear and often lacking
- Misalignment between IT solutions and research questions/needs/situations (e.g. **problems with access to software**)
- **No sustainable plans for maintenance and updates of most data infrastructures**
- No sustainable plans for tracking and accessing related samples/materials

DATA_SCIENCE
datastudies.eu

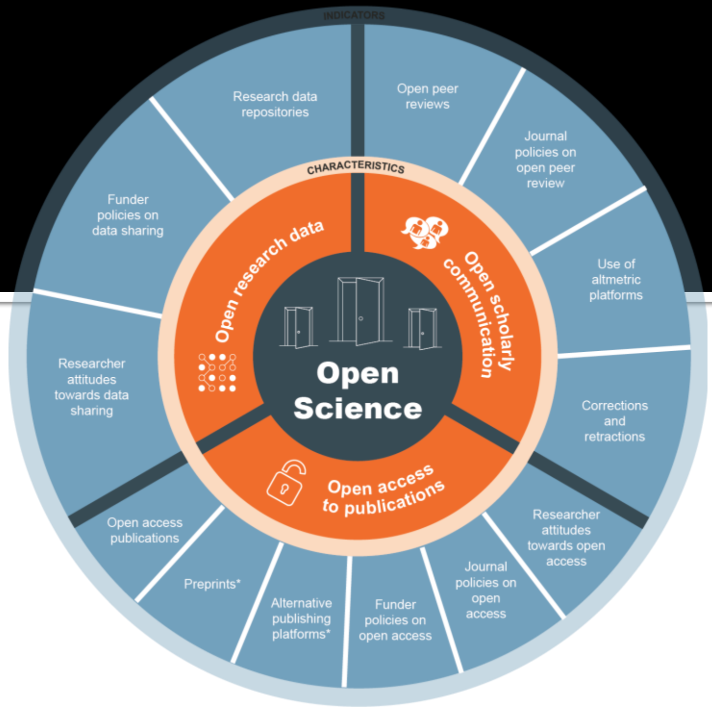
egenis

Research funded by **European Research Council** grant n° 335925, **ESRC** grant ES/P011489/1 , **ESRC**, **MRC** & **NERC** MEDMI Grant, **Leverhulme Trust** Grant “Beyond the Digital Divide”, **Australian Research Council** Discovery Grant “Organisms and Us”



French translation **“La recherche scientifique à l’ère des Big Data. Cinq façons dont les données nuisent à la science, et comment la sauver »** available from **April 2019**

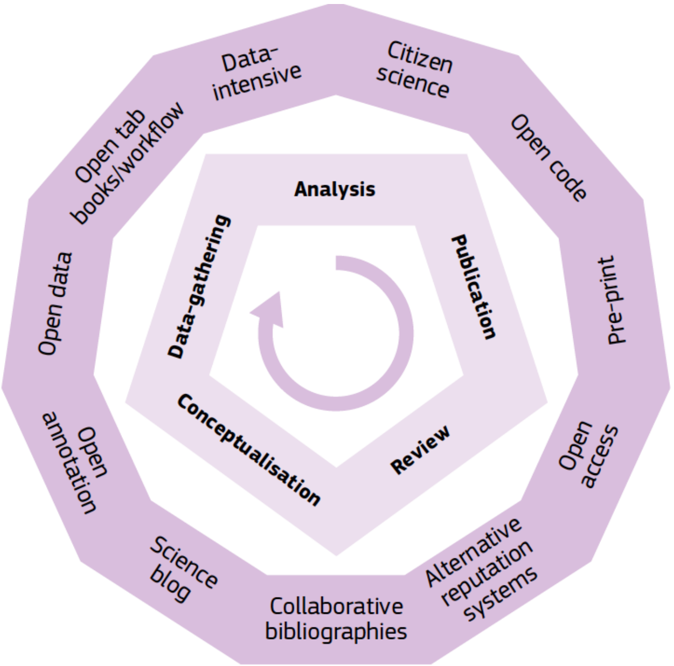
Open Research



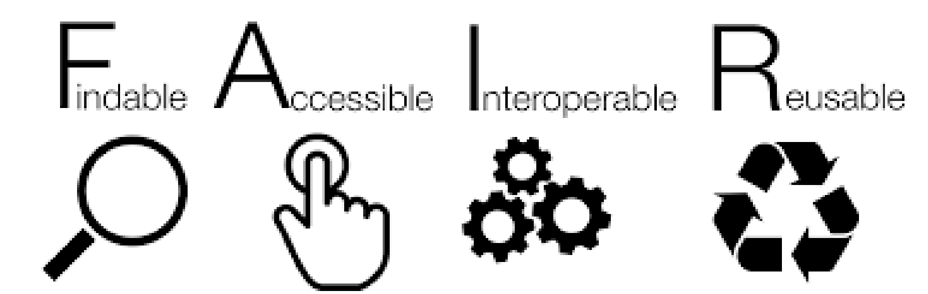
Variously defined by

- the use of new digital tools
- a specific set of values
- practices of collaboration and sharing
- a view of the research workflow and related governance

Platform to debate what counts as science, scientific infrastructures and scientific governance, and how results should be credited and disseminated



The Value of FAIR Data



Potential to improve

- pathways to and quality of discoveries
- uptake of new technologies
- collaborative efforts across disciplines, nations and expertises
- research evaluation, debate and transparency
- appropriate valuation of research components beyond papers and patents
- fight against fraud, low quality and duplication of efforts
- legitimacy of science and public trust
- public understanding and participation

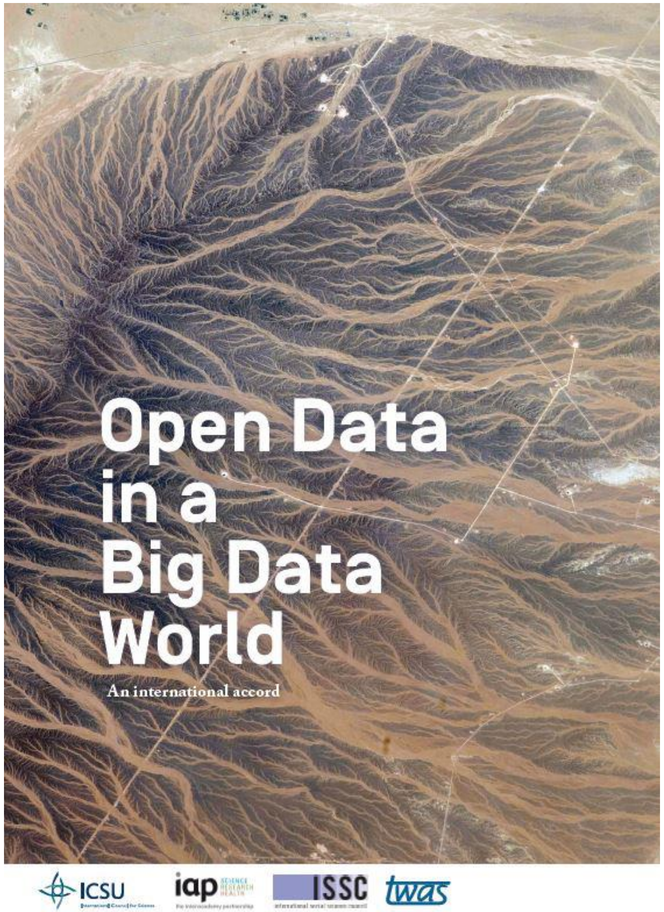
Open Research

Widespread agreement on three aspects:

- GLOBAL SCOPE: affects all stages of the research process, its implementation involves a wide set of governance structures

- SYSTEMIC REACH: involves systemic shift in current practices of research, publishing and evaluation

- LOCAL IMPLEMENTATION: its implications for any one research systems need to be considered with reference to its specific characteristics -- thus the mechanisms through which OS is implemented are likely to vary



Abstract

This talk discusses the conditions under which Open Data can be effectively disseminated, mined and reused so as to be fruitful to research and provide a platform for new discoveries. For Open Data to benefit research, considerable resources need to be invested in the developing strategies and tools that facilitate data sharing, as well as in assessing and regularly re-evaluating the scientific, social, cultural and economic implications of such strategies. I demonstrate this through an examination of the history and current characteristics of existing practices of data management and re-use across the biological and biomedical sciences. I focus specifically on the study of 'data journeys', that is the ways in which data are made to travel beyond the sites in which they were originally produced. Such study reveals several key challenges for Open Science implementation, which I discuss in detail. I shall conclude that adequate, labour-intensive data curation is crucial to tackling these challenges in ways that are reliable and sustainable in the long term.

Lesson Learnt 1: Context-Specific Curation Is Key to Data Reuse

- Pluralism in methods and standards contributes robustness to data analysis, and reduces risk of losing system-specific knowledge

- **Interoperability is preferable to integration**
  - Standards ad formats are key
  - Yet reliance on overly rigid standards creates exclusions and obliterates system-specific knowledge
  - Data linkage methods are best when it is possible to disaggregate

IITA

Transforming African Agriculture

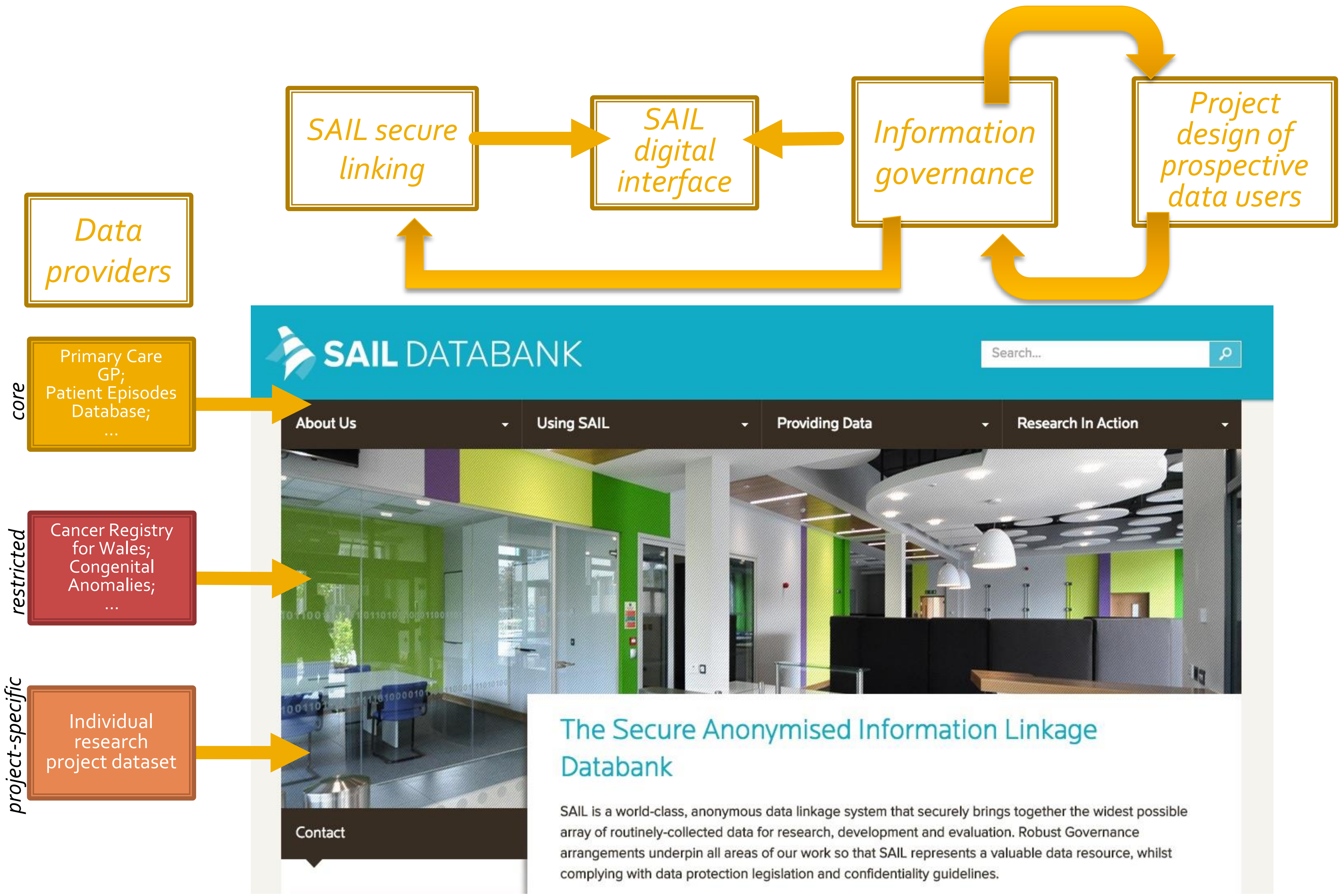
50

1967 - 2017

CGIAR

RESEARCH PROGRAM ON

Roots, Tubers and Bananas


SAIL secure linking
SAIL digital interface
Information governance
Project design of prospective data users
Data providers
core
Primary Care GP; Patient Episodes Database; …
restricted
Cancer Registry for Wales; Congenital Anomalies; …
project-specific
Individual research project dataset
SAIL DATABANK
Search...
About Us
Using SAIL
Providing Data
Research In Action
Contact
The Secure Anonymised Information Linkage Databank
SAIL is a world-class, anonymous data linkage system that securely brings together the widest possible array of routinely-collected data for research, development and evaluation. Robust Governance arrangements underpin all areas of our work so that SAIL represents a valuable data resource, whilst complying with data protection legislation and confidentiality guidelines.