An ELIXIR Perspective

Carole Goble
ELIXIR UK Head of Node
ELIXIR Interoperability Platform co-lead

Jo McEntyre
ELIXIR EMBL-EBI
ELIXIR Data Platform co-lead

www.elixir-europe.org

DCIP Publisher Early Adopters workshop, 22 July 2016, London.
ELIXIR connects national centres and EMBL-EBI to build a sustainable European infrastructure for biological research data.

ELIXIR underpins life science research – across academia and industry.
major bioinformatics service providers (~150)

20 ELIXIR members
2 observers

Co-operation

Long term support
Organisation in a nutshell

- Data
- Tools
- Interoperability (Standards)
- Compute
- Training
Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplary implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Fortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of ‘long-term care’ of valuable digital assets, with the goal that they should be discovered and reused for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes ‘good data management’ is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guidelines to inform those who publish and preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessible, Interoperable, Reusability—that serve to guide data producers and publishers as they navigate these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure the research process must be transparent, reproducible, and reusable.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other’s data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: bened.mens@dit.ni).

A full list of authors and their affiliations appears at the end of the paper.
Identification and analysis of the CtNUDIX gene. In a previous study, we constructed an in planta BNS-specific cDNA plasmid library from the susceptible Canadian lentil cultivar Eston infected with C. truncatum isolate CT-21 and identified 122 unigenes encoding proteins with putative signal peptides, including effectors (17). In this study, we characterized one of the effector proteins, Ct21-1373, now named CtNUDIX.

The CtNUDIX cDNA (GenBank accession no. [GB] HO663661) is 1,184-bp (bp) in size with an ORF of 669 bp, which encodes a 223 aa protein. An SP of 24 aa with a cleavage site between alanine-24

Gene Expression Studies

DOIs ('long tail')

Inherited disease(OMIM)

DNA Variations (SNPs)

Protein Data Bank

European Nucleotide Archive

Kaftas S, Kim JH, and McEntyre JR Database Citation in Full Text Articles (May 2013) PLoS One 10.1371/journal.pone.0063184
Data Citation

1. Impact of data and data resources
   – Evidence to select, support and sustain infrastructure
   – “Indicators” of community usage
   – Cited use of resource

2. Europe PubMedCentral
   – Core ELIXIR data resources
   – Integration of literature with data key to inclusive and effective infrastructure
   – Data citation (and consequently bidirectional linking)

3. Curation & Identifier Services & Practices
   – Joined up services for identifiers, citation and credit
   – CDL/EBI identifier harmonisation
   – identifiers.org, n2t.net, ezid, datacite, orcid ...
   – Drive practices, including data curation workflows

4. Dataset metadata
   – Standards, practices, indexers, catalogers, tools, adoption
   – Scaled up finding and citation using Search Engines
Indicator: “Community served”

Usage
• IP addresses/sessions on web site per month for past 2/3 years
• Page/data requests for web site, FTP, web services per month for past 2/3 years

Use of resource in research
• No. times the resource mentioned in research articles per year (in Europe PMC)
• No. times accession numbers from resource mentioned or cited in research articles (in Europe PMC)
• Key “database” papers (e.g. published in NAR Database issue) and the number of citations.

Dependency
• on the resource by others service (what is the reach through)?
Cataloguing and Indexing Datasets (and their content)

BioSchemas: Exploitation of schema.org

Partnership:
• ELIXIR
• NIH BD2K
• Google
Bonus Slide
Citation: G. Penkler; F. du Toit; W. Adams; M. Rautenbach; D. C. Palm; D. D. van Niekerk; J. L. Snoep; (2014): Glucose metabolism in Plasmodium falciparum trophozoites; FAIRDOMHub. http://dx.doi.org/10.15490/seek.1.investigation.56
ELIXIR: http://www.elixir-europe.org/

Bioschemas: http://www.bioschemas.org

NIH BD2K bioCADDIE
• https://biocaddie.org/
• DATS: https://biocaddie.org/workgroup-3-group-links
• DATAMED: https://datamed.org/
• https://biocaddie.org/datamed-prototype-call-feedback
Links

ELIXIR: http://www.elixir-europe.org/

Bioschemas: http://www.bioschemas.org

NIH BD2K bioCADDIE
• https://biocaddie.org/
• DATS: https://biocaddie.org/workgroup-3-group-links
• DATAMED: https://datamed.org/
• https://biocaddie.org/datamed-prototype-call-feedback

FAIRDOM: http://www.fair-dom.org
Research Objects: http://www.researchobject.org