Institution: Heriot-Watt University

Unit of Assessment: B11 Computer Science

Title of case study: Making research data more findable, accessible, interoperable, and reusable

Period when the underpinning research was undertaken: 2013 – 2015

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Alasdair Gray	Associate Professor	Sep 2013 – present

Period when the claimed impact occurred: 2014 – 2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact

Data has been an underutilised output of research projects for many years due to the challenges of finding, understanding, and reusing that data. Research at Heriot-Watt Computer Science by Dr Gray contributed substantially to the definition of the FAIR Data Principles (2016) and led to a global Health Care and Life Sciences community recommendation for describing datasets for discovery and reuse. This community recommendation has been adopted by several data providers including the RDF Platform of the European Bioinformatics Institute (EBI). The standard has also been adopted and used internally in major pharmaceutical companies, including AstraZeneca, leading to datasets that comply with the FAIR Data Principles to be more readily reused and exploited.

2. Underpinning research

From February 2011 until May 2015, the World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLS-IG) had an activity, co-led by Dr Gray as an invited expert, to develop a community profile for describing datasets, i.e. to provide machine readable metadata to make datasets more Findable, Accessible, Interoperable, and Reusable, cf. the FAIR Data Principles [3.1].

The group first identified use cases from a broad range of applications within the Health Care and the Life Sciences domains. One such use case was drawn from Dr Gray's work on the Open PHACTS Data Platform [3.2] where there was a requirement to know which version of a dataset was used within the platform, and to identify in query responses where each item of data had been retrieved from.

A community profile that extended the Data Catalog Vocabulary (DCAT) was developed to meet the needs of the use cases. Dr Gray extended the core model of DCAT to support the abstract notion of a dataset distinct from versions and distributions of the dataset. The capability developed by Dr Gray enabled the model not just to reference a dataset but to attach the

information on all of its version history. As an example, it enables references to the dataset ChEMBL or any of the specific versions or multiple distribution formats per version [3.3]. The community also agreed on which properties were to be mandatory, recommended, and optional, and detailed statistics needed to enable data reuse. The community profile was published as a W3C Interest Group Note in May 2015 [3.4] and subsequently described in [3.5].

To support the adoption of the community profile, Dr Gray exploited his earlier experiences of adoption in the Open PHACTS project to provide concrete examples from a real-world dataset that could easily be adapted for other datasets. Additionally, Dr Gray's team at HWU developed a validation tool, deployed by the W3C (https://www.w3.org/2015/03/ShExValidata/) that supported verifying the correctness of dataset descriptions against the profile (Hansen et al., 2015). The tool supports users providing their dataset description and then choosing the level of conformance to validate against. A crucial aspect of the tool was providing meaningful, contextualised, error messages when a dataset description deviated from the community profile. Dr Gray with his collaborator Prof Dumontier fed the outcomes of the HCLS community profile work into the development of the FAIR Data Principles [3.1] in particular shaping principles F1-3, A1 and A2, I1-3, and R1 and its sub-clauses. Together with a wider team of collaborators we demonstrated the ability to make data available following the FAIR Data Principles and the W3C HCLS Community Profile [3.6].

3. References to the research

[3.1] Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, CA, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, M, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J & Mons, B 2016, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, vol. 3, 160018. https://doi.org/10.1038/sdata.2016.18

[3.2] Groth, P, Loizou, A, Gray, AJG, Goble, C, Harland, L & Pettifer, S 2014, 'API-centric linked data integration: the Open PHACTS discovery platform case study', *Journal of Biomedical Semantics*, vol. 29, pp. 12-18. https://doi.org/10.1016/j.websem.2014.03.003

[3.3] Hansen, JB, Beveridge, A, Farmer, R, Gehrmann, L, Gray, AJG, Khutan, S, Robertson, T & Val, J 2015, 'Validata: An online tool for testing RDF data conformance', Paper presented at 8th International Conference on Semantic Web Applications and Tools for Life Sciences 2015, Cambridge, United Kingdom, 7/12/15 - 10/12/15.

[3.4] Gray, AJG (ed.), Baran, J (ed.), Marshall, MS (ed.), Dumontier, M (ed.), Alexiev, V, Ansell, P, Bader, G, Bando, A, Bolleman, JT, Callahan, A, Cruz-Toledo, J, Gaudet, P, Gombocz, EA, Gonzalez-Beltran, A, Groth, P, Haendel, M, Ito, M, Jupp, S, Juty, N, Katayama, T, Kobayashi, N, Krishnaswami, K, Laibe, C, Le Novère, N, Lin, S, Malone, J, Miller, M, Mungall, CJ, Rietveld, L, Wimalaratne, SM & Yamaguchi, A 2015, *Dataset Descriptions: HCLS Community Profile*. W3C Interest Group Note, World Wide Web Consortium. <https://www.w3.org/TR/hcls-dataset/>

[3.5] Dumontier, M, Gray, AJG, Marshall, MS, Alexiev, V, Ansell, P, Bader, G, Baran, J, Bolleman, JT, Callahan, A, Cruz-Toledo, J, Gaudet, P, Gombocz, EA, Gonzalez-Beltran, A, Groth, P, Haendel, M, Ito, M, Jupp, S, Juty, N, Katayama, T, Kobayashi, N, Krishnaswami, K, Laibe, C, Le Novère, N, Lin, S, Malone, J, Miller, M, Mungall, CJ, Rietveld, L, Wimalaratne, SM & Yamaguchi, A 2016, 'The health care and life sciences community profile for dataset descriptions', *PeerJ*, vol. 4, e2331. https://doi.org/10.7717/peerj.2331

[3.6] Wilkinson, MD, Verborgh, R, Olavo Bonino da Silva Santos, L, Clark, T, Swertz, MA, Kelpin, FDL, Gray, AJG, Schultes, EA, van Mulligen, EM, Ciccarese, P, Kuzniar, A, Gavai, A, Thompson, M, Kaliyaperumal, R, Bolleman, JT & Dumontier, M 2017, 'Interoperability and FAIRness through a novel combination of Web technologies', *PeerJ Computer Science*, vol. 3, e110. https://doi.org/10.7717/peerj-cs.110

4. Details of the impact

The increasing use of computers to support researchers in gathering data, processing and analysing data, and publishing data and research results has led to a step-change in the way research is conducted. It was vitally important to pharmaceutical companies that the Open PHACTS system could provide provenance on the returned query answers, to state where the data originated (ChEMBL, Drugbank, UniProt, etc) and which version of the dataset was used. Dr Gray developed the Open PHACTS Dataset Description to provide the needed metadata about the data consumed, including important properties such as stating the version and format of the ingested data. This allowed specific provenance information to be returned on the platform's query answers and increased trust in the analysis resulting from the data.

Building on the Open PHACTS Dataset Descriptions, subsequent research in the period 2013-2015 led to The FAIR Data Principles, published in March 2016, which set out desirable criteria to enable the discovery, retrieval, understanding, and reuse of data associated with research, particular that funded by public bodies such as UK Research & Innovation (UKRI), European Research Council (ERC), and National Institutes of Health (NIH) [5.1].

The FAIR Data Principles built on Dr Gray's work on dataset descriptions, particularly with respect to the definitions of principles F1, F2, F3, A1, A2, I1, I3, R1.1, R1.2, R1.3. Dr Gray collaborated in the development of the FAIR Principles and engaged in activities to publicise the FAIR principles and train people to FAIRifiy their data. The FAIR Data Principles were endorsed by the G20 Leaders' Communique Hangzhou Summit, September 2016, by stating;

"We support effort to promote voluntary knowledge diffusion and technology transfer on mutually agreed terms and conditions. Consistent with this approach, we support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles". [5.2]

The Principles have subsequently led to interest within industry and academia to further exploit data that has previously been collected, either internally by companies or publicly by academia. This has been particularly the case within the health care and life sciences community where pharmaceutical companies have initiated/funded initiatives to retroactively make existing datasets comply with the FAIR Data Principles so that they can be more readily reused and exploited.

The W3C HCLS Dataset Description Profile enables the meeting of FAIR principle R1.3 and has been adopted internally in major pharmaceutical companies, including AstraZeneca as a means to make their internal data more discoverable and reusable by a wider set of research labs across the world. AstraZeneca's Director, Oncology Translational Medicine, Data Strategy Lead stated,

"We recognised the costs associated with continual curation and reshaping of data as new questions arise beyond the original collection intent and have found the alignment and implementation of the FAIR principles as a way to solve this challenge", and, "we found the use of the DCAT standard and the W3C HCLC recommendations to be critical to implementing the FAIR data set management". [5.3]

The profile has been deployed within major data repositories including the European Bioinformatics Institute's (EBI) RDF Platform, the Swiss Institute for Bioinformatics (SIB), and the Japanese RIKEN MetaDatabase portal for life sciences data. At the EBI, the profile was used to automate their data ingestion pipeline for their RDF platform. The approach allowed them to perform various quality control checks on the metadata. This improved the quality of the data and also saved time [5.4]

An independent study recognised that Biopharma Research and Development (R&D) productivity can be improved by implementing the FAIR Data Principles and is an enabler for digital transformation of Biopharma R&D. The study went on to highlight the impact for one company who had implemented a FAIR platform for 3,000 users across three main sites and, "*since running the FAIR platform for 2 months, the company collected usage activity data based on click counts per user. This FAIR platform had 900,000 page views in 60 days. The projection for the year gave an estimation of ~5.5M page views. A very conservative assumption that each of these FAIR-enhanced views saved ~5 s, by providing better search results with direct access to the target repository, led to a calculation of ~3.5 full-time employees (FTEs) worth of time saved per year*" [5.5].

In Boston, 14 May 2020, The Pistoia Alliance, a global, not-for-profit alliance that works to lower barriers to innovation in life sciences R&D, launched a freely accessible toolkit to help companies implement the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for data management and stewardship. Collated by experts in the field, the toolkit contains numerous method tools, training and change management, as well as use cases, allowing organizations to learn from industry successes. The Alliance recognised that as the life sciences industry continues to digitize, the FAIR guiding principles of *Findable, Accessible, Interoperable* and *Reusable* data would help organizations realise their digital transformation.

"*At Roche, we know that implementing the FAIR principles can be difficult for biotech and pharma organizations of every size, so we are very pleased to lead on this project and help make the process easier,"* commented the Principal Scientist at Roche. *"The toolkit will help to smooth the path to greater data sharing within and between industries, which is critical to future research efforts. We see the FAIR guiding principles as a worthy goal, and one which will help the industry realize the value of technologies like deep learning."* [5.6]

5. Sources to corroborate the impact

[5.1] National Institute of Health NIH New Models of Data Stewardship – endorsement of FAIR Data Principles https://commonfund.nih.gov/commons/awardees.

[5.2] G20 Leaders' Communique Hangzhou Summit, September 2016, point 12 – endorsing the FAIR Data Principles were endorsed
https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967

[5.3] Letter from AstraZeneca's Director, Oncology Translational Medicine, Data Strategy Lead – confirming use of FAIR Data Principles.

[5.4] Head of Molecular Archival Resources, European Bioinformatics Institute (EMBL-EBI), who can be contacted to confirm implementation and impact of adopting the FAIR Data Principles.

[5.5] John Wise., Alexandra Grebe de Barron., Andrea Splendiani., Beeta Balali-Mood., Drashtti Vasant., Eric Little., Gaspare Mellino., Ian Harrow., Ian Smith., Jan Taubert., Kees van Bochove., Martin Romacker,, Peter Walgemoed,, Rafael C. Jimenez., Rainer Winnenburg., Tom Plasterer., Vibhor Gupta., Victoria Hedley., 2019. Implementation and relevance of FAIR data principles in biopharmaceutical R&D, Drug Discovery Today, Volume 24, Issue 4, https://doi.org/10.1016/j.drudis.2019.01.008

[5.6] The Pistoia Alliance announcement – Launch of toolkit to accelerate implementation of FAIR Data Principles https://www.pistoiaalliance.org/news/fair-toolkit-launch/