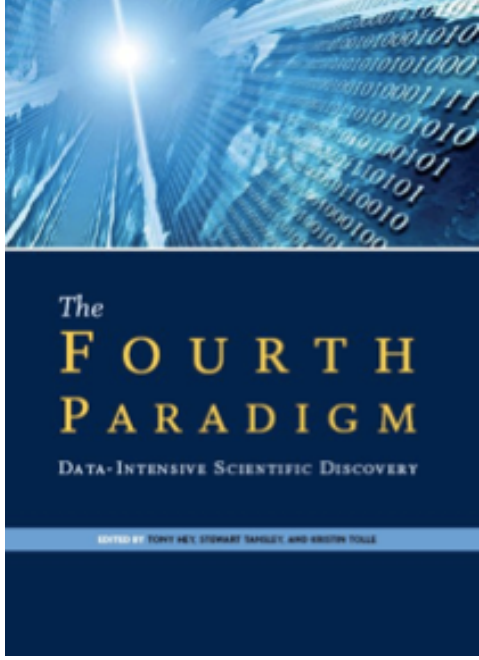
Philosophy of data-intensive science

Sabina Leonelli
Department of Sociology, Philosophy and Anthropology & Egenis
University of Exeter

Data-intensive science: A new paradigm?



- New technologies for the production, storage and dissemination of data: computing power is seen as transforming how science is done, but no coherent and systematic assessment of such transformation to date

- How is science changing to take advantage of digital technologies for data dissemination? With which implications?

- Long history of data collection and sharing in science: what is new today, and how do these practices differ from other forms of scientific inquiry?

- What can and cannot be learnt from 'big data', and how? Can science be 'data-driven'? How can the quality, relevance and reliability of data be assessed?

Possible epistemic drawbacks: fertile terrain for philosophical investigation

- Information **overload** versus interpretation and synthesis: what are we actually learning from 'big data'?
- Issues with standards: can they be **trusted**? **Who** develops them and how?
- Danger of **conservatism** (available data are favoured)
- Issues with **quality** controls and **peer review** (burden for peers, unclear status of data in the process, reproducibility)
- New opportunities for **fraud** (bad data, digital manipulation of evidence, plagiarism)

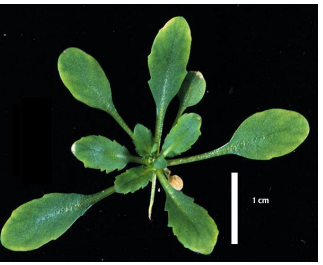
How are data actually disseminated and re-used? Focus on *data journeys*

• Understanding how data are actually circulated and used is key to understanding what counts as scientific knowledge in the digital era, including what counts as evidence, theory and experiment

• Data travel requires work, including significant conceptual and material scaffolding that then affects further research; need for intelligent ways to make data 'open' (Royal Society 2012)

• Understanding contexts / domains in which data acquire evidential value is crucial

• Hence: focus on use of online databases to make data travel



My Empirical Work

Methods - ***Empirically grounded philosophy of science***: following the data, archival research, interviews, policy engagement on 'open science' and ***collaboration*** with curator and user communities

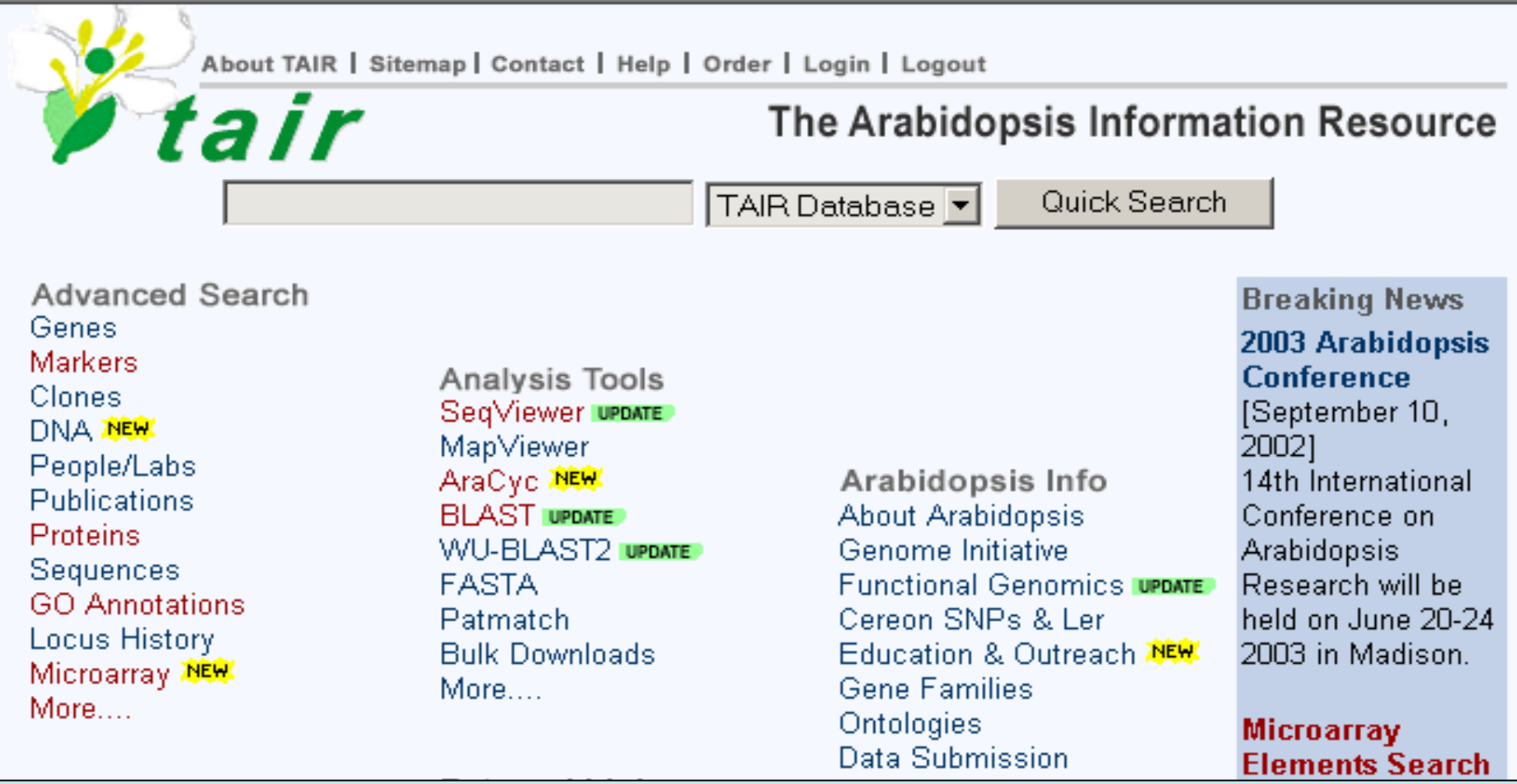
Focus - Model organism research: bringing together various types of data on the 'same' organism [e.g. 'community databases']; increasingly serving also cross-species and translational research

- Leonelli, S. (2013) Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science. *Studies in the History and Philosophy of the Biological and Biomedical Sciences.*
- Leonelli, S. and Ankeny, R.A. (2012) Re-Thinking Organisms: The Epistemic Impact of Databases on Model Organism Biology. *Studies in the History and Philosophy of the Biological and Biomedical Sciences.*
- Leonelli, S. (2010) Packaging Data for Re-Use: Databases in Model Organism Biology. In Howlett, P. and Morgan, M.S. (eds) *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge University Press

Key difficulty in these areas: pluralism

(no centralisation of experiments and data formats)

Model Organism Databases: Defining Standards for Collection, Dissemination and Interpretation of Data on Organisms

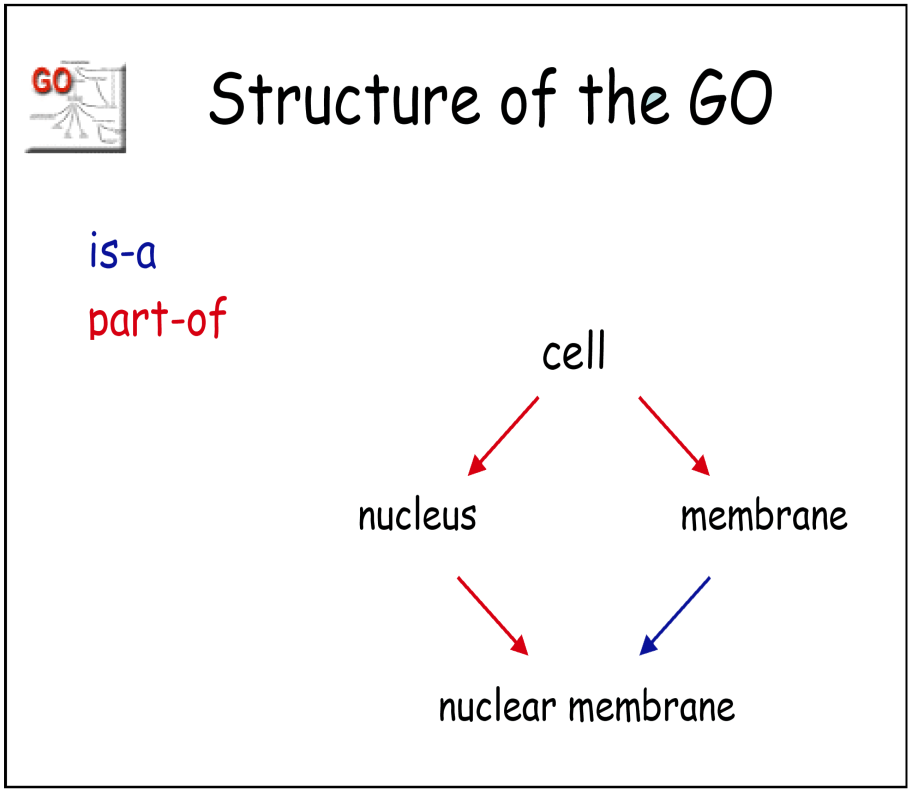


"our goal is to provide the common vocabulary, visualisation tools, and information retrieval mechanisms that permit integration of all knowledge about *Arabidopsis* into a seamless whole that can be queried from any perspective'

The Gene Ontology

'formal representations of areas of knowledge in which the essential *terms* are combined with structuring rules that describe the *relationship* between the terms. Knowledge that is structured in a bio-ontology can then be *linked to the molecular databases*'

- Precisely defined, *descriptive terms*
- Precisely defined *relations* among terms
- Association of terms with *datasets*



Result: network of interdependent claims about phenomena

Transforming data into knowledge: Stages of data journeys

(1) De-contextualisation: making data travel across research contexts

[Temporary] separation of data from information about their provenance.

This requires adequate standards and guidelines for data formatting.

(2) **Re-contextualisation: assessing data quality and reliability**

Meta-data: adding information about provenance enables re-contextualisation of data production

Efficient meta-data presuppose reliable reference to material specimens (e.g. strains in stock centres), experimental protocols, instruments and calibration techniques

(3) Re-use: using data towards discovery

No simple induction / 'automated reasoning': data interpretation involves reference to theories embedded in specific practices

Results

What counts as ‘good data’ in model organism biology?

- Depends on experimental standards
- Serious disagreements and diversity across subfields

Data classification as a theory-making activity (e.g. bio-ontologies)

Understanding of data re-use (feeding into policy discussions of Open Science)

Reconceptualising the organism

Reconceptualising knowledge production: Comparison of alternative ways to organise data is key to further understanding and exploration of significance of data

Questioning the reach of the ‘Fourth Paradigm’

Key Publications

- Leonelli, S. (accepted) Data Interpretation in the Digital Age. *Perspectives on Science*.
- Leonelli, S. (2013) Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science. *Studies in the History and Philosophy of the Biological and Biomedical Sciences: Part C.* Online First.
- Leonelli, S. (2012) Classificatory Theory in Biology. *Biological Theory*, 7(1). Online First.
- Leonelli, S. (2012) Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies*. International Studies in the Philosophy of Science* 26(1): 47-65.
- Leonelli, S. (2012) When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42(2): 214-236.
- Leonelli, S. (2012) Making Sense of Data-Driven Research in the Biological and the Biomedical Sciences. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 43(1): 1-3.
- Leonelli, S. and Ankeny, R.A. (2012) Re-Thinking Organisms: The Epistemic Impact of Databases on Model Organism Biology. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 43(1): 29-36.
- Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. and Lomax, J. (2011) How the Gene Ontology Evolves. *BMC Bioinformatics*, 12:325 (tagged 'highly accessed').
- Ankeny, R.A. and Leonelli, S. (2011) Bioethics Authorship in Context: How Trends in Biomedicine Challenge Bioethics. *The American Journal of Bioethics*, 11(10): 22-24.
- Bastow, R. and Leonelli, S. (2010) Sustainable digital infrastructure. *EMBO Reports,* 11(10): 730-735.
- Leonelli, S. (2010) Machine Science: The Human Side. *Science,* 330 (6002): 317.
- Leonelli, S. (2010) Documenting the Emergence of Bio-Ontologies: Or, Why Researching Bioinformatics Requires HPSSB. *History and Philosophy of the Life Sciences*, 32, 1: 105-126.
- Leonelli, S. (2010) Packaging Data for Re-Use: Databases in Model Organism Biology. In Howlett, P. and Morgan, M.S. (eds) How Well Do Facts Travel? The Dissemination of Reliable Knowledge. Cambridge University Press, pp.325-348.
- Leonelli, S. (2010) The Commodification of Knowledge Exchange: Governing the Circulation of Biological Data. In: Radder, H. (ed) The Commodification of Academic Research: Science and the Modern University. Pittsburgh UP, pp.132-157.
- Leonelli, S. (2009) Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia. In Atkinson, P., Glasner, P. and Lock, M. (eds) The Handbook for Genetics and Society: Mapping the New Genomic Era*.* Routledge, pp. 469-485.
- Leonelli, S. (2009) On the Locality of Data and Claims About Phenomena. *Philosophy of Science,* 76, 5: 737-749.
- Leonelli, S. (2008) Bio-Ontologies as Tools for Integration in Biology. *Biological Theory,* 3, 1: 8-11.