Optimising Scientific Knowledge Transfer: On Collective Sensemaking, And the Radical Idea

Anita de Waard, VP Research Data Collaborations, Elsevier
http://www.slideshare.net/anitawaard/
350 years ago, science was an individual enterprise:

An Extract of a Letter, containing some Observations, made in the ordering of Silk-worms, communicated by that known Vertuolo, Mr. Dudley Palmer, from

Of a Way of killing Rattle-Snakes.

A Spot in one of the Belts of Jupiter.

The Ingenious Mr. Hook did, some months since, intimate to a friend of his, that he had, with an excellent twelve foot Telescope, observed, some days before, he than spoke of it, (videl. on the ninth of May, 1664. about 9 of the clock at night) a small Spot in the biggest of the 3 obscurer Belts of Jupiter, and that, observing it from time to time, he found, that within 2 hours after, the said Spot had moved from East to West, about half the length of the Diameter of Jupiter.
Today, science is done by big groups:
But that he was much surprised when he saw the Micrography of Mr. Hook, and found there, that his engine was published on a mere Theory, without having made any Experiment, though that might have been made with little charge and great speed; expense of Money and Time being the only thing that can excuse those inns to the public without make trial thereof.

In the rest of this note, we show that while these are some beautiful ideas, the above analysis hides some important subtleties.

where $n_s$ is the scalar spectral index, $r$ is the tensor-to-scalar ratio, and $N_e$ is the number of e-foldings of inflation (usually $50 \lesssim N_e \lesssim 60$). We will discuss these predictions further in Section 7.
How do we unify the needs of the collective and the individual?

Goal: Support the development of radical new ideas:
- Allowing *attribution to claims and evidence*,
- Anchored in a *substrate of integrated data*,
- Collectively create *nimble and robust* systems of knowledge management
- *For all citizens!*

“Let us endeavor to build systems that allow a kid in Mali who wants to learn about proteomics to not be overwhelmed by the irrelevant and the untrue.”

- John Perry Barlow, iAnnotate 2014
1. Introduction

The Standard Model (SM) of particle physics [1], [2], [3] and [4] has been tested by many experiments over the last four decades and has been shown to successfully describe nature. However, the mechanism that breaks electroweak symmetry experimentally. This mechanism [5], [6], [7], [8], [9] and [10] implies the existence of a scalar particle, the Higgs, the only elementary particle in the SM that has not yet been observed at the Hadron Collider [11] (LHC) physics programme. citations are to papers, not thoughts

Both the ATLAS and CMS Collaborations reported excesses of events in their 2011 datasets of proton-proton (pp) collisions at centre-of-mass energy $\sqrt{s} = 7$ TeV at the LHC, which were compatible with SM Higgs boson production and decay in the mass region 124–126 GeV with 8.8 standard deviations ($\sigma$), respectively [17] and [18]. The CMS Collaboration recently reported a broad excess in the mass region 120–150 GeV with observed local significances for $m_H = 125$ GeV are 2.6 and 2.5 in the 2011 and 2012 data combinations [16].

Data is presented as pictures, not - data

Fig. 2. The $C_L$ values for the SM Higgs boson hypothesis as a function of the Higgs boson mass in the range 110–600 GeV (left) and 110–145 GeV (right). The observed values are shown by the solid line. The dashed line indicates the expected median of results for the background-only hypothesis, while the green (dark) and yellow (light) bands indicate the ranges that are expected to contain 68% and 95% of all observed excursions from the median, respectively. The three horizontal lines on the $C_L$ plot show confidence levels of 90%, 95%, and 99% defined as $(1 - C_L)$. For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.
Better ways of linking:
Today: Linking Data to Journals

- Supplementary data at PANGAEA
- Bidirectional links between PANGAEA & ScienceDirect
- Data visualized next to the article

http://www.elsevier.com/databaselinking
Better ways of linking: Tomorrow: All-Publishers Linking Service

- ICSU/WDS/RDA Publishing Data Service Working group
- Currently creating linked-data model for exposing DOI to DOI links outside publisher’s firewall
- Merged with National Data Service pilot with the same goal
- Collaboration between CrossRef, DataCite, Europe PubMed Central, ANDS, Thompson Reuters, Elsevier...

**Objective:** move from a plethora of (mostly) bilateral arrangements between the different players...

.. to ..

.. a one-for-all cross-referencing service for articles and data
A second problem: data cannot be found – or integrated...

Data storage systems are not aimed at knowledge integration....

... and manual curation cannot keep up!
A Maslow Hierarchy for Data:

1. Preserved (existing in some form)
2. Archived (long-term & format-independent)
3. Accessible (can be accessed by others)
4. Comprehensible (others can understand data & processes)
5. Discoverable (can be indexed by a system)
6. Reproducible (others can redo experiments)
7. Trusted (validated/checked by reviewers)
8. Citable (able to point & track citations)
9. Usable (allow tools to run on it)
Better Models of Data Integration:
Today: A Semantic Model for the Research Object

- Work done by Matt Brush at OHSU & colleagues:
- Overview of semantic models for ‘Research Object’
- Use to base data integration projects within/outside of Elsevier on
- Next: use as basis for shared model for DDI/Force11 working group

Two Components of a Research Object Metadata Model
Better Models of Data Integration:
Tomorrow: Data FAIRPort, Mark Wilkinson

- **Work done** by Mark Wilkinson @ U Madrid for FAIR (Findable, Accessible, Interoperable, Reusable) Project
- Enable access to different repositories through common schema without requiring effort from underlying repositories:
A third problem: scientific software is not up to snuff...

The tools we use to explore scientific/technical knowledge are greatly outdated...

...while the tools we use our daily lives are getting smarter all the time.
Better systems for building scientific software

Today: RDA Cost Recovery WG

• Cochair with Ingrid Dillo (DANS), Simon Hodson (CODATA)
• Goal: write a report regarding new potential funding models for data repositories, allow them to start sharing this knowledge
• Interviewed 24 repositories on their funding (current and future)
• Now summarising stories and trends – looking for help!

Terms of funding for main income stream (in %)
Better systems for building scientific software

Tomorrow: Conglomerates of Partners

• **Force11**:  
  – Multi-stakeholder, member-driven organisation  
  – Unites scholars, tool developers, **librarians**, publishers, funding agencies etc. etc.  
  – Works through working groups: Data Citation, RII, now: Divini, Software citation (with W3C, NDS and Github)  
  – Made bid for community component of NIH Commons  
  – Conference in Portland with OHSU (Melissa Haendel, Robin Champieux): **April 17-19, 2016**

• **National Data Service**:  
  – Multi-stakeholder group, based around supercomputing centres  
  – Aims to be a ‘connective tissue’ between data creation, curation, storage etc projects.  
  – Inviting Pilots: two or more partners who have not worked together, interested in collaborating on a data-centric project to solve a real-world need  
  – Next Consortium meeting: October (19-21?) 2015– here!  
  – **NEED MORE INVOLVEMENT FROM LIBRARY COMMUNITY!**
In summary:

Let’s collectively enable ‘an account of the present undertakings, studies and labours of the ingenious in many considerable parts of the world’, by connecting thoughts and data, through superior systems and software!

Anita de Waard
VP Research Data Collaborations, Elsevier
a.dewaard@elsevier.com
@anitadewaard

Let’s keep talking!
Appendix: Recommendations

<table>
<thead>
<tr>
<th>Goal</th>
<th>Requirement</th>
<th>Action</th>
<th>Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>To see the heritage of a particular idea:</td>
<td>Need better tools to cite claims/evidence, not papers</td>
<td>Build tools</td>
<td>Software creators</td>
</tr>
<tr>
<td></td>
<td>Need requirements to trace ideas – not papers</td>
<td>Issue mandates, recommendations</td>
<td>Funders, societies, journals?</td>
</tr>
<tr>
<td>To improve how data is integrated:</td>
<td>Need systems/tools to find and integrate data, methods</td>
<td>Build connections, search engines</td>
<td>Software creators</td>
</tr>
<tr>
<td></td>
<td>Need standards, habits, requirements, mandates to post data, methods, analyses</td>
<td>Mandates, standards</td>
<td>Funding bodies, learned societies, journals</td>
</tr>
<tr>
<td>To develop better tools for scientific software:</td>
<td>Need robust funding models for sustainable software development</td>
<td>Develop models for long-term development</td>
<td>Funding agencies, governments, i.c.w. industry if needed</td>
</tr>
<tr>
<td></td>
<td>Need better metrics for assessing software quality, usability, search, etc.</td>
<td>Develop metrics for scientific software performance</td>
<td>Software creators, funding agencies, all</td>
</tr>
</tbody>
</table>