Making it count
A computational approach to attribution

A project to make more meaningful connections between people, their roles, their work and impacts.

29 October 2018
Kristi Holmes, PhD
#RO2018. @kristiholmes
Who we are and who we serve

The community we serve

>50 CTSA Hubs

9 CD2H Sites

NIH National Center for Advancing Translational Sciences

The larger informatics community
**CD2H: National Center for Data to Health**

Data & Informatics Coordinating Center for the CTSA Program

**Accelerating Informatics Innovation to Advance Translational Research**

- Make Data Easier to Share and Re-use
- Make Tools More Accessible and Interoperable
- Leverage Expertise and Foster a More Collaborative CTSA Culture

Better translation of research and improved patient care
What *IS* impact?

More than papers and grants – we are driving toward improved health and wellbeing

- Improvements in health through treatment and prevention
- Contributions to society through economic growth and productivity
- Expansion of the biomedical knowledge base through cutting-edge research
- Cultivation of the biomedical workforce of today and tomorrow

For effective translation of knowledge and discoveries into the improved health of our communities, it is essential to incorporate evaluation strategies that enable investigators and teams to measure, monitor, and communicate the impact of their work

https://www.nih.gov/about-nih/what-we-do/impact-nih-research
### Contributors and expertise needed for a genetic diagnosis

<table>
<thead>
<tr>
<th>Clinical/care</th>
<th>Pathology</th>
<th>Ontologist</th>
<th>CS/informatics</th>
<th>Curator</th>
<th>Basic research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Markello</td>
<td>Hans Goebel</td>
<td>Melissa Haendel</td>
<td>Elizabeth Lee</td>
<td>Melissa Haendel</td>
<td>Johannes Grosse</td>
</tr>
<tr>
<td>Dong Chen</td>
<td></td>
<td></td>
<td>Amanda Links</td>
<td>David Adams</td>
<td>Attila Braun</td>
</tr>
<tr>
<td>Justin Y. Kwan</td>
<td></td>
<td></td>
<td>Will Bone</td>
<td>David Draper</td>
<td>David Varga-Szabo</td>
</tr>
<tr>
<td>Iren Horkayne-Szakaly</td>
<td></td>
<td></td>
<td>Murat Sincan</td>
<td>Bailey Gallinger</td>
<td>Niklas Beyersdorf</td>
</tr>
<tr>
<td>Alan Morrison</td>
<td></td>
<td></td>
<td>Damian Smedley</td>
<td>Joie Davis</td>
<td>Boris Schneider</td>
</tr>
<tr>
<td>Olga Simakova</td>
<td></td>
<td></td>
<td>Jules Jacobson</td>
<td>Nicole Vasilevsky</td>
<td>Lutz Zeitlmann</td>
</tr>
<tr>
<td>Irina Maric</td>
<td></td>
<td></td>
<td>Nicole Washington</td>
<td>Heather Trang</td>
<td>Petra Hanke</td>
</tr>
<tr>
<td>Jay Lozier</td>
<td></td>
<td></td>
<td>Elise Flynn</td>
<td>Rena Godfrey</td>
<td>Patricia Schropp</td>
</tr>
<tr>
<td>Andrew R. Cullinane</td>
<td></td>
<td></td>
<td>Sebastian Kohler</td>
<td>Gretchen Golas</td>
<td>Silke Mühlstedt</td>
</tr>
<tr>
<td>Tatjana Kilo</td>
<td></td>
<td></td>
<td>Orion Buske</td>
<td>Catherine Groden</td>
<td>Carolin Zorn</td>
</tr>
<tr>
<td>Lynn Meister</td>
<td></td>
<td></td>
<td>Marta Girdea</td>
<td>Michele Nehrebeckey</td>
<td>Michael Huber</td>
</tr>
<tr>
<td>Kourosh Pakzad</td>
<td></td>
<td></td>
<td>Michael Brudno</td>
<td>Ariane Soldatos</td>
<td>Carolin Schimmittwolf</td>
</tr>
<tr>
<td>Sanjay Chainani</td>
<td></td>
<td></td>
<td>Jeremy Band</td>
<td>Elise Valkanas,</td>
<td>Wolfgang Jagla</td>
</tr>
<tr>
<td>Roxanne Fischer</td>
<td></td>
<td></td>
<td></td>
<td>Colleen Wahl</td>
<td>Philipp Yu</td>
</tr>
<tr>
<td>Camilo Toro</td>
<td></td>
<td></td>
<td></td>
<td>Lynne Wolfe</td>
<td>Thomas Kerkau</td>
</tr>
<tr>
<td>James G. White</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Harald Schulze</td>
</tr>
<tr>
<td>David Adams</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Michael Nehls</td>
</tr>
<tr>
<td>Cornelius Boerkoel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bernhard Nieswandt</td>
</tr>
<tr>
<td>William A. Gahl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cynthia J. Tifft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meral Gunay-Aygun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

@ontowonka
Contributors and expertise needed for a genetic diagnosis

Clinical/care  Pathology  Ontologist  CS/informatics  Curator  Basic research

How can we credit all of these contributors?

Moreover, how can we find the resources and people to form the scientific teams, collaborations, reviewers, we need?

A community journey

Contributor Roles & Research Outputs
Better attribution: extending credit beyond the publication

What work is being done, who is doing it, and what outputs are being created?

1. Understand deeply the requirements for a computable attribution system from a large diversity of stakeholders;
2. Build model(s) to meet these requirements (CRO, ROO);
3. Evaluate the models in real pilot systems with real data.

By using contribution roles & research outputs to develop infrastructure to understand the scholarly ecosystem, we can better understand, leverage, and credit a diverse translational workforce.
The Informatics of Attribution

1. Understand deeply the requirements for a computable attribution system from a large diversity of stakeholders;
2. Build model(s) to meet these requirements; and
3. Evaluate the models in real pilot systems with real data.

Development of data models to address these needs demands a rigorous requirements-driven approach

Key modeling challenges for development of integrative community standards

1. Accommodation of diverse and complex data types
2. Support needs of different applications and systems
3. Interoperability with broader data landscape

Key tools necessary to drive change

1. Technology
2. Persistent identifiers
3. Data models
4. Connections – of all kinds!

@kristiholmes & @ontowanka
1. Provide a VIVO experience for everyone, a demonstration of VIVO, a platform for experimentation, and an ownership experience for the VIVO team

2. Use persistent identifiers for all entities – people (ORCiD), works (DOI and PMID), organizations (GRID), journals (ISSN), concepts (FAST)

3. Automatic, real-time ingest of metadata from identifiers via public APIs

4. Publication of data

5. Consumption and reuse of data

6. Attribution of works by scholars to indicate roles in works

--Mike Conlon, VIVO Project Director

My own profile was completed entirely with publicly available data via ORCID and DOIs and it took about 15 minutes to complete from start to finish.

http://openvivo.org/
1. Provide a VIVO experience for everyone, a demonstration of VIVO, a platform for experimentation, and an ownership experience for the VIVO team.

2. Use persistent identifiers for all entities – people (ORCiD), works (DOI and PMID), organizations (GRID), journals (ISSN), concepts (FAST).

3. Automatic, real-time ingest of metadata from identifiers via public APIs.

4. Publication of data.

5. Consumption and reuse of data.

6. Attribution of works by scholars to indicate roles in works.

--Mike Conlon, VIVO Project Director
OpenVIVO: Transparency in Scholarship

Violeta Ilk, Michael Conlon, Graham Triggs, Marijane White, Muhammad Javed, Matthew Brush, Karen Gutzman, Shahim Essaid, Paul Friedman, Simon Porter, Martin Szomszor, Melissa Anne Haendel, David Eichmann and Kristi L. Holmes

1 Stony Brook University, Stony Brook, NY, United States
2 University of Florida, Gainesville, FL, United States
3 RAND, CA, United States
4 University of Chicago, IL, United States
5 Digital Science, London, United Kingdom
6 University of Iowa, Iowa, IA, United States

OpenVIVO is a free and open-hosted semantic web platform that anyone can join and that gathers and shares open data about scholarship in the world. OpenVIVO, based on the VIVO open-source platform, provides transparent access to data about the scholarly work of its participants. OpenVIVO demonstrates the use of persistent identifiers, the automatic real-time ingest of scholarly ecosystem metadata, the use of VIVO-ISF and related ontologies, the attribution of work, and the publication and reuse of data—all critical components of presenting, preserving, and tracking scholarship. The system was created by a cross-institutional team over the course of 3 months. The team created and used RDF models for research organizations in the world based on Digital Science GRID data, for academic journals based on data from CrossRef and the US National Library of Medicine, and created a new model for attribution of scholarly work. All models, data, and software are available in open repositories.

Transparency in Scholarship

Scholarship requires knowledge of previous work. The growth of scholarship worldwide and the proliferation of scholarly output types—from papers and monographs to preprints, conference papers, datasets, posters, and presentation slides—have fundamentally changed the scholarly ecosystem from an environment dependent on libraries to one that is dependent on the electronic resources made available by libraries to support discovery and knowledge transfer. This shift clearly drives a need for the representation of scholarly works using standard metadata formats to facilitate indexing and discovery.

For scholars to have knowledge of previous work, the work must be indexed and discoverable via electronic systems. Metadata regarding the scholarly work is important for the discovery of the work, but there is a need for explicit transparency and citation practices. OpenVIVO addresses this need by providing a platform that allows for transparent and persistent representation of scholarly work. OpenVIVO allows researchers to create persistent identifiers for their work, which can be used to track and attribute the work. The platform also provides tools for the attribution of work, including the use of VIVO-ISF and related ontologies. This allows for a more transparent and comprehensive representation of scholarly work, which is essential for the advancement of research and scholarship.
OpenVIVO: Transparency in Scholarship

Join us! Submit tickets! 😊
https://github.com/data2health/contributor-role-ontology


Transparency in Scholarship

Scholarship requires knowledge of previous work. The growth of scholarship worldwide has produced an ecosystem from an environment dependent on libraries to one that is dependent on the electronic discovery and knowledge transfer. This shift clearly drives a need for the representation of scholarly work. All models, data, and software are accessible.
CRediT is high-level taxonomy, including 14 roles, that can be used to represent the roles typically played by contributors to scientific scholarly output. The roles describe each contributor’s specific contribution to the scholarly output.

CRediT ontology in OWL:
https://github.com/data2health/credit-ontology

Nature 508, 312–313 (17 April 2014) doi:10.1038/508312a

https://casrai.org/credit/
It takes technology + culture.
Why now & how do we get there?

Informatics of Attribution

- Translational workforce
- Team Science
  - Open Science
- Diverse research activities & outputs
- Greater drive:
  - Analysis
  - Accountability
  - Advocacy
  - Allocation
- Stakeholder collaboration & engagement
- Advances in research information systems
- Big Data: empowerment & opportunity

Research Objects: A Common Unit of Sharing Across Use Cases

- Computable Workflows Comprised of Reusable Content
- Repositories of Reusable Content (Data, Information, Knowledge)
- "Signals" of Research Activity and Knowledge Sharing

Extensible, standards-based “primitive” that defines a minimal shareable product (MSP)

Reproducible and rigorous methodologies
Data Sets, Software, Formalized Knowledge, Publications
Publication, Contribution, Attribution, Dissemination
Teams
- CD2H
  - **NU team**: Karen Gutzman, Patty Smith, Sara Gonzales
  - **OHSU team**: Marijane White, Nicole Vasilevsky, Melissa Haendel
- Northwestern University Clinical and Translational Sciences Institute
- OpenVIVO collaborators, Force11 Attribution WG, NISO, Cathy Sarli & Becker Library
- Galter Library, NUCATS, ChicagoCHEC, FIRST DailyLife, Health for All

NIH Support
- U24TR002306 (NCATS)
- UL1TR001422 (NCATS)
- U54CA202995, U54CA202997, U54CA203000 (NCI)
- P30AR072579 (NIAMS)
- G08LM012688 (NLM)

Contact us!
kristi.holmes@northwestern.edu
@kristiholmes
https://ctsa.ncats.nih.gov/cd2h/
@data2health
Links to selected resources and projects

• National Center for Advancing Translational Sciences: [https://ncats.nih.gov/](https://ncats.nih.gov/)
• Clinical and Translational Science Award (CTSA) Program: [https://ctsacentral.org/](https://ctsacentral.org/)
• Northwestern University Clinical and Translational Sciences Institute: [https://nucats.northwestern.edu/](https://nucats.northwestern.edu/)
• CD2H: [https://ctsa.ncats.nih.gov/cd2h/](https://ctsa.ncats.nih.gov/cd2h/) and [https://github.com/data2health](https://github.com/data2health)
• FORCE11 Attribution Working Group: [https://www.force11.org/group/attributionwg](https://www.force11.org/group/attributionwg)