Actionable Outcomes

Developed from the Belmont Forum e-Infrastructures and Data Management Curricula Development Workshop
28 April, Vienna, Austria

Premise

Encourage that data should be

- Discoverable, accessible, understandable, managed and protected
- Managed in accordance with e-I&DM data management outline recommendations

The primary audience for training/curricula are researchers in data-intensive global change research, especially those pursuing inter- or trans-disciplinary research.

Training provision and attracting attendees

1. Training should be outcomes-based (rather than time spent); therefore assessment should be considered as part of delivery of training activities.
2. Any training provided should be practical, interactive and useful in data and situations that a research scientist would encounter.
3. Training should remain current and relevant as techniques evolve and improve.
4. Data from other domains could be used in training to allow for a collaborative interactive event e.g. hackathon – which could be linked to particular research challenges such as those which arise from the Sustainable Development Goals.
5. Workshops on specific e-I&DM topics could be held as side activities to large conferences or as part of other domain science activities, such as summer schools. Inserting small modules of data management topics into large scientific domain activities would be a good way to capture a large and diverse attendance.
6. Training should focus on the benefits of improving data management, workflow etc. by including tangible incentives, e.g. increasing citation by using persistent identifiers and citing precision. Useful practical exercise: can you find your own data?
7. Match training options with the appropriate project role. For example, a PI with limited time may simply be briefed on data principles and transdisciplinary approaches, whilst a postdoc who will be intimately involved in data production, use and analysis may have more comprehensive multi-day training on workflow, programming and all aspects of data management.
8. The training may be delivered by data professionals or domain scientists with excellent data management awareness and skills.

The hybrid model

9. Awareness of a range of digital skills is important – if only to know what you don't know. It is important to acknowledge that a role exists of a "hybrid scientist," which incorporates skills of both a computer scientist and domain scientist. Any domain scientist may be on the spectrum of a hybrid (depending on other team skills available). The hybrid (or digital scientist) may write libraries or perform data transformations, and can be used as a target

model for capacity building.

10. There is a recognised role for data professionals as contributors to optimal data outcomes throughout the life of a project; needs may range from consulting with institutional or repository data managers to more active and involved participation. Thus, a training priority is to facilitate communication between researchers and data professionals, to determine when and where a data professional is called for and to designate trainers with 'local' knowledge of data management practices.

Certification

11. The field of e-I&DM is fast moving, and its swift rate of change may be a barrier to providing both a list of recommended courses and certification. Recognition of the value of a course by the general research and data management community is more important and current than certification.
12. There could be an advantage for both learners and training providers for a certificate being issued by the Belmont Forum as endorsement of a course within the specified curriculum. This certification should be used as an endorsement and not an accreditation.
13. A certificate could be issued jointly with a university or learned society if this improved the value or appeal of the certificate.
14. Courses may be presented as a variety of badges based on core competencies. Belmont Forum researchers may undertake a self assessment questionnaire to determine which skills they already have and which badges they need.

Curricula content

15. Good data and metadata standards are key to improving many skills, and increased awareness of how to create and arrange metadata would help scientists ask the right questions when they encounter new datasets and reduce apparent 'data complexity'. Also included is the idea of awareness of paradata, which includes provenance.
16. Important to train on standardising:
    - 16.1. Semantic vocabularies across domains
    - 16.2. Syntactic ways to encode data
    - 16.3. Data and software citation
17. Knowledge of the full research data life cycle including data management and planning, data collection and workflows, and long-term preservation is vital – this will provide context for research scientists and help design projects with research outcomes that can be shared and reused.
    - 17.1. Peer-reviewed best practice for data, including ethical and transparency issues;
    - 17.2. Journal and funder data policies;
    - 17.3. How to find relevant data and reporting standards for a specific discipline;
    - 17.4. Data 101: basic training on spreadsheet and table formatting, file naming, README file creation
    - 17.5. Documenting data workflows

The role of the Belmont Forum

18. Require Belmont Forum funded project scientists to attend certain training activities, for example data management.

19. The Belmont Forum is in a powerful position to provide thought leadership in the form of position papers. Relevant topics include:
    19.1. The importance of open data and examples of how research has been improved with open data along with open enterprise examples.
    19.2. Encouragement of basic knowledge of the research data lifecycle (including the role of trustworthy data repositories) and best practice in data management.
    19.3. Recommending standard vocabularies and encoding as appropriate; including persistent identifiers.
    19.4. Showcasing and endorsing researchers who promote and carry out data sharing, especially to other scientific domains.
    19.5. The need for academic institutions and publishers to recognize and value open research and collaboration for hiring, promotion, and publication.
20. The Belmont Forum can target a funding call on the re-use of data, especially involving discovery of data outside one's domain. There could be conditions in the call on collaboration and exchange of staff as secondments.

Observations

- Many courses already exist on the topics discussed, along with organisations with websites that collate lists of courses. Existing training resources include open online courses, ESIP DMT Clearinghouse as a registry for data management courses, Software or Data Carpentry for programming and The Kepler Project for scientific workflow. Therefore, availability of appropriate material appears secondary, and the primary challenge may be to encourage (through funding) researchers to work differently and integrate data management into their work. Steps are needed to make it normal for research scientists to be conscientious and open about their data. To persuade researchers to adopt actions surrounding data management, which are not seen as one's core activity, the activity needs other promotion, e.g. compulsion or incentives, until data management actions become accepted practice.
- There are many stakeholders in the research data lifecycle, and researchers should not be the only ones responsible. Data repositories, data managers and academic institutions need to provide supporting roles and appropriate resources.
- To be able to provide annotation or ranking of currently available training resources would be beneficial to the community, to identify preferred activities and courses.
- Terminology and definitions are important both for accuracy of data use and for appealing to the right audience when titling training. Use 'data organisation' rather than workflow and 'data assessment' rather than data complexity.
- Data repositories will increasingly become data publishers.
- The biomedical profession/researchers are expected to communicate their data – this culture needs to be engendered in other sciences.
- Transdisciplinary principles of stakeholder driven research orientated towards action should be considered as part of delivery of training activities.
- In addition to training, an ongoing relationship between data professionals and researchers would be lead to better data archiving outcomes.