The Fourth Paradigm 10 Years On

Tony Hey · Anne Trefethen

Introduction

The book, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, was a collection of provocative, forward-looking essays published in 2009. It now seems a good time to look back at some of the significant developments in data-intensive science and scholarly publishing that have happened in the last 10 years and see how the predictions of the authors have fared.

The book was dedicated to Turing Award winner Jim Gray of Microsoft Research, who was tragically lost at sea in January 2007. Jim's vision for this Fourth Paradigm for research had its origins nearly 10 years earlier with Jim in the USA, and with "eScience" in the UK. In the late 1990s Jim had recognized that the next "Big Data" challenge for database technologies would likely come from science rather than from commerce. He understood the very real technical challenges that the management and analysis of very large scientific datasets would pose for scientists, and the key role that IT and computer science could play in extracting new science from their data. In the UK, the Director General for Research, John Taylor, had initiated an "eScience" programme in 2001 to help meet the challenge of the coming era of data-intensive science. The eScience programme covered many scientific research fields and was primarily focused on the technologies needed to manage, analyze, visualize and curate "Big Scientific Data". Or in Jim Gray's words, "eScience is where IT meets scientists".

The Fourth Paradigm: visions and reality

The Fourth Paradigm book contains many intriguing insights and predictions. We note some from each section below, together with a brief commentary on how these projections compare with the situation in scientific research 10 years on.

Earth and environment

- **From Jeff Dozier and Bill Gail on "The Emerging Science of Environmental Applications"**
  The emerging third phase, knowledge developed primarily for the purpose of scientific understanding is being complemented by knowledge created to target practical decisions and action. This new knowledge endeavor can be referred to as the ***science of environmental applications****.*
- **From Jim Hunt, Dennis Baldocchi and Catharine van Ingen on "Redefining Ecological Science Using Data"**
  These changes require a new approach to resolving resource management questions. … Addressing these challenges requires a synthesis of data and models that span length scales from the very local (river pools) to the global (oceanic circulations) and spans time scales from a few tens of milliseconds to centuries.

https://doi.org/10.1007/s00287-019-01215-9

Tony Hey
Rutherford Appleton Laboratory,
Science and Technology Facilities Council,
Didcot OX11 0QX, United Kingdom
E-Mail: Tony.Hey@stfc.ac.uk

Anne Trefethen
Oxford e-Research Centre, Department of Engineering Science,
University of Oxford,
Oxford OX1 3PJ, United Kingdom

Alle "Aktuellen Schlagwörter" seit 1988 finden Sie unter:
http://www.is.informatik.uni-wuerzburg.de/as

– **From John Delaney and Roger Barga on "A 2020 Vision for Ocean Science"**
The cabled ocean observatory merges dramatic technological advancements in sensor technologies, robotic systems, high-speed communications, eco-genomics, and nanotechnology with ocean observatory infrastructure in ways that will substantially transform the approaches that scientists, educators, technologists, and policymakers take in interacting with the dynamic global ocean.

Commentary. Each day the National Oceanic and Atmospheric Administration (NOAA) collects over 20 terabytes of data and **data.noaa.gov** hosts over 97 thousand data sets [1]. This is just one of the sources of data about the ocean and the life within in it. As Delaney and Barga predicted, the availability of such data is now beginning to transform both policy and behaviour. This does not necessarily happen without some prompting. The Big Ocean Button Challenge, for example, offered prizes for apps based on using this data to provide services for fishing, shipping, ocean acidification, public safety and exploration [2].

As another indication that attitudes to data are changing, funding agencies now insist that all research proposals contain a data management plan. In addition, several digital data repositories have emerged for storing research data where no funded, discipline-based archive is available. In the US, the Dryad repository for research data curation and publishing was established in 2012: in 2018 there were almost 45,000 downloads from Dryad's 24,000 datasets [3]. In Europe, the European Commission's OpenAIRE project partnered with CERN to set up Zenodo[1], a general repository for European funded research outputs, both data and software [4]. In July 2019, the Alfred P. Sloan Foundation funded a partnership between Dryad and Zenodo "*to make open research practices more seamless for researchers*" [5].

Health and well-being

– **From Michael Gillam et al. on "The Healthcare Singularity and the Age of Semantic Medicine"**
Today, the dissemination path for medical information is complex and multi-faceted, involving commercials, lectures, brochures, colleagues, and journals. In a world with nearly instantaneous knowledge translation, dissemination paths would become almost entirely digital and direct.
– **Horvitz and Kristan: Toward a Computational Microscope for Neurobiology**
We foresee that neurobiologists studying populations of neurons will one day rely on tools that serve as ***computational microscopes*** *– systems that harness machine learning, reasoning, and visualization to help neuroscientists formulate and test hypotheses from data. Inferences derived from the spatiotemporal data streaming from a preparation might even be overlaid on top of traditional optical views during experiments, augmenting those views with annotations that can help with the direction of the investigation.*
– **Buchan, Winn, and Bishop: A Unified Modeling Approach to Data-Intensive Healthcare**
We anticipate a fourth paradigm of healthcare information … whereby an individual's health data are aggregated from multiple sources and attached to a unified model of that person's health. The sources can range from body area network sensors to clinical expert oversight and interpretation, with an individual playing a much greater part than at present in building and acting on his or her health information. Incorporating all of this data, the unified model will take on the role of a "health avatar" – the electronic representation of an individual's health as directly measured or inferred by statistical models or clinicians. Clinicians interacting with a patient's avatar can achieve a more integrated view of different specialist treatment plans than they do with care records alone.

Commentary. These insights in the potential for healthcare and related areas to be transformed by new forms of data and by aggregating different data sources is now the major challenge for national healthcare systems. However, although much progress has been made we are still far from the vision of a "healthcare singularity" in which medical knowledge flows frictionlessly and immediately from research to practice. Similarly, although Horvitz and Kristan's vision for a computational microscope for neuroscience has not been fully realized, it is clear that machine learning technologies are becoming increasingly important for use with healthcare data to predict healthcare outcomes.

[1] Zenodo is derived from Zenodotus, the first librarian of the Ancient Library of Alexandria and father of the first recorded use of metadata, a landmark in library history.

For example, researchers at the European Molecular Biology Laboratory (EMBL) have developed computational methods that allow the analysis of multiple types of molecular data from individuals. Such a multi-omics approach integrates genomic, epigenomic, transcriptomic, metabolomic and other molecular data to build a profile of a given patient. The Multi-Omics Factor Analysis (MOFA) designed by the EMBL team has been tested on data collected from leukemia patients and shown to lead to improved diagnosis. This is a pre-cursor for the personalized treatment of cancer and other diseases [6].

The amounts of data available in the form of clinical and pathological images as well as patient biometric data are now leading to the beginnings of truly personalized medicine. In fact, machine learning technologies have now been shown to have *better than human performance* for certain tasks such as image recognition [7] and strategy games [8]. AI algorithms are now being used in areas such as the detection of tumours and melanomas, where they have been shown to be able to differentiate between images of malignant and benign skin lesions as well as certified dermatologists [9]. There are many other examples of the growing realization of digital healthcare and of machine learning algorithms that support the vision of the authors of the Fourth Paradigm [10, 11].

Research Infrastructure

- **Alex Szalay and Jose Blakeley on "Gray's Laws: Database-centric Computing in Science"**
  Cloud computing is a recently emerging paradigm. It offers obvious advantages, such as co-locating data with computations and an economy of scale in hosting services.
- **Mark Abbott on "A New Path for Science"**
  Today, semantic web and ontologies are being proposed as a means to enable knowledge discovery and collaboration. However, as with databases, it is likely that the science community will be reluctant to use these inherently complex tools except for the most mundane tasks.
- **Christopher Southan and Graham Cameron on "Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data"**
  ELIXIR is now a reality. … *the mission of the ELIXIR project … aims to ensure a reliable distributed infrastructure to maximize access to biological information that is currently distributed in more than 500 databases throughout Europe.*
- **Carole Goble and David De Roure on "The Impact of Workflow Tools on Data-Centric Research"**
  … *data-centric science could be characterized as being about the primacy of data as opposed to the primacy of the academic paper or document, but it brings with it a method deluge: workflows illustrate* ***primacy of method*** *as another crucial paradigm in data-centric research.*

Commentary. Cloud computing is now a reality, and in addition to commercial offerings from Amazon, Microsoft and Google and others, we are now seeing the emergence of "on-premise" Cloud infrastructure and hybrid clouds, which connect these on-premise computing resources to commercial clouds [12].

The last decade has also seen a move towards recognizing research infrastructure as an important component of the national infrastructure with data resources now also classified as a legitimate part of such a national infrastructure. In the UK, the National Infrastructure Commission [13] has commissioned a number of activities under the premise of data for the public good [14]. This recognizes data as infrastructure [15] and focuses on the collection of the right data and on standards for sharing data – with data both from government as well as government research agencies. The creation of the important initiative for *Data and Analytics for National Infrastructure* (DAFNI) is recognition of the importance of both data resources and computational data needs [16].

The data infrastructure to support research has evolved hugely since 2009, with many nations taking forward open data initiatives and building the infrastructure and tooling to support these aims. The Australian site **data.gov.au** is one of the better examples of these developments in which data are now being treated as a national resource that allows easy access and search capabilities. Other countries have followed suit, and the UK and USA each have open data sites **data.gov.uk** and **data.gov**, respectively, that provide similar access to national data resources.

These developments certainly support the primary importance of data. However, in addition to data, the importance of method has matured through the further developments in scientific workflows. A special edition of Future Generation

Computer Systems [17] follows developments of workflow systems over the last decade. Workflow systems have clearly advanced substantially in ease of use, support for improved abstractions, automation of data identification, and inbuilt tools for provenance. The different workflow communities came together in 2014 to propose the Common Workflow Language, CWL [18]. This is an open standard for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments. It is designed to meet the needs of data-intensive science, such as bioinformatics, medical imaging, astronomy, physics and chemistry. The ELIXIR project is now taking a leading role supporting experiments with CWL workflows, as well as many other services for the life science communities [19].

Scholarly Communication

- **Clifford Lynch on "Jim Gray's Fourth Paradigm and the Construction of the Scientific Record"**
  With the arrival of the data-intensive computing paradigm, the scientific record and the supporting system of communication and publication have reached a Janus moment when we are looking both backward and forward. It has become clear that data and software must be integral parts of the record – a set of first-class objects that require systematic management and curation in their own right.
- **Paul Ginsparg on "Text in a Data-centric World"**
  So we should neither overestimate the role of data nor underestimate that of text, and all scientists should track the semantic enhancements of text and related data-driven developments in the biological and life sciences with great interest – and perhaps with envy.
- **Herbert van der Sompel and Carl Lagoze on "All Aboard: Towards a Machine-Friendly Scholarly Communication System"**
  Recently, we have witnessed a significant push toward a machine-actionable representation of knowledge embedded in the life sciences literature, which supports reasoning across disciplinary boundaries. Advanced text analysis techniques are being used to extract entities and entity relations from the existing literature, and shared ontologies have been introduced to achieve uniform knowledge representation. This approach has already led to new discoveries based on information embedded in the literature that was previously readable only by humans.

Commentary. Recognition of the importance of publishing data, either alongside journal publications, or as a dataset in its own right, has grown enormously in the last decade. In the UK, most university research repositories now include both full texts of research papers and supporting research data. At the University of Oxford, for example, the Oxford University Research Archive, ORA, now supports the submission of data as well as articles and other submissions [20]. The Nature Science Data journal is celebrating its 5th birthday this year [21, 22]; the journal includes data publication, best practice, standards and related aspects. It is indicative of the changes that have taken place in science infrastructure and the important place of data.

The global movement towards "Open Science" and "research reproducibility" have also played an important role in establishing research data and software as first-class objects. The OECD defines Open Science as making: "*the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction*" [23]. The EC FOSTER project – Fostering the Practical Implementation of Open Science in Horizon 2020 and Beyond [24] – proposes that open science should be more than just at the basic level of the OECD definition. In their view, open science is about **extending the principles of openness to the whole of the research cycle**, fostering sharing and collaboration as early as possible – a principle that harks back to the vision of the UK eScience initiative.

What has not been so successful is the predicted widespread take-up of semantic web technologies – such as ontologies, RDF, OWL and SPARQL – much beyond the biological sciences research community. It appears that Mark Abbott's rather more pragmatic analysis of the use of these technologies may be the reality for most scientific fields. The recent emergence of bioschemas as an extension to schema.org could be a more practical way forward to adding useful semantic information to both data and documents [25, 26]. In addition, the use of JSON-LD for manipulating linked data has proved to be easier and more accessible to a wider technical audience than the original semantic web technologies [27].

Major developments since 2009

AI and the deep learning revolution. What other major developments were not strongly identified in 2009? The first and most obvious technical omission is the "deep learning" technology pioneered by Geoffrey Hinton, Yann LeCun and Joshua Bengio. These three were the recipients of the 2019 Turing Award [28], and the award citation is "for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing".

A key starting point for the Deep Learning revolution we are now witnessing dates back to the ImageNet database and the AlexNet deep learning network [29]. ImageNet was a project led by Professor Fei-Fei Li from Stanford University and produced a database containing annotations for over 14 million high-resolution images available on the Web. The images were labeled by human labelers recruited using Amazon's Mechanical Turk. Starting in 2010, a competition called the ImageNet Large-Scale Visual Recognition Challenge was held using the database. The competition used a subset of the ImageNet collection with roughly 1000 images in each of the 1000 categories. In all, there were roughly 1.2 million training images, 50,000 validation images and 150,000 testing images. The intent was to provide the computer science community with a focus for evaluating the effectiveness and progress of computer vision systems. A landmark breakthrough in image classification was made in the 2012 competition by Geoffrey Hinton and two of his PhD students, Alex Krizhevsky and Ilya Sutskever. AlexNet, as their neural network implementation came to be called, used a "deep neural network" consisting of five convolutional layers and three fully connected layers and was implemented using two GPUs. Their paper won the 2012 ImageNet competition and reduced the error rate by an astonishing 10.8 % compared to the previous winner [30]. The 2015 competition was won by a team from Microsoft Research using a deep neural network with over 100 layers and achieved an error rate for object recognition comparable to human error rates [31]. In the words of Geoffrey Hinton, the "deep learning is an algorithm which has no theoretical limitations on what it can learn; the more data you give and the more computational time you provide, the better it is" [32].

Can such AI and deep learning algorithms benefit scientific research? Google's DeepMind subsidiary in the UK has brought together physicists, machine learning experts and structural biologists to create a system called "AlphaFold" [33, 34]. The DeepMind team entered the biennial competition organized by CASP (critical assessment of protein structure prediction) that assesses the state of the art in three-dimensional protein structure modeling. The predictions of the AlphaFold system were remarkably good and better on average than the other 97 competitors.

Towards open science: the OSTP memorandum, Plan-S and the FAIR principles. In February 2013, the US Office of Science and Technology Policy in the Executive Office of the President issued a memorandum requiring that Federal agencies investing in research develop clear policies to support increased public access to the results of their research [35]. The memo stipulated that "such results include peer-reviewed publications and digital data", where digital data is defined as:

the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.

This memorandum was soon followed by similar declarations from the Global Research Council in May [36] and from the G8 Science Ministers in June 2013 [37].

All the major US Federal research funding agencies have now developed their policies for "increased public access" of the research that they fund. This includes open access to research papers and the need for researchers to have serious data management plans in their proposals. Since US researchers funded by these agencies contribute a large fraction of all US research papers, there is clearly increasing global momentum towards "open science". This necessarily requires not only open access to research publications but also to the metadata and data required to validate and make sense of the research results.

More recently in Europe, in 2018 Plan-S was proposed as an open access initiative in Europe [38]. The plan is supported by cOAlition S, an interna-

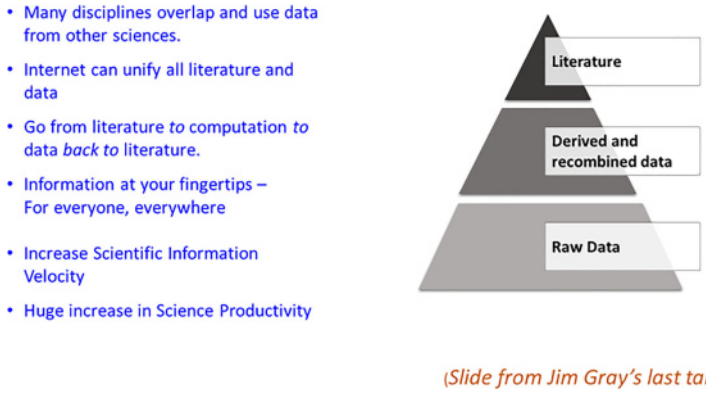

Fig. 1 Jim Gray's Vision: All Scientific Data Online

tional consortium of research funders and has the aspiration that from 2021, scientific publications that result from research funded by public grants must be published in compliant Open Access journals or platforms.

The FAIR data principles published in Scientific Data in 2016 are likely to play an important role in this area [39]. This proposes guidelines to make digital assets more *F*indable, *A*ccessible, *I*nteroperable, and *R*eusable. The principles emphasize machine-actionability – defined as the capability of computational systems to find, access, interoperate, and reuse data with little or no human intervention. This is necessary as humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data. Alongside the production of these standards and policies, sites have been developed to curate data and metadata standards, inter-related to databases and data policies [40, 41].

Concluding remarks

The example of AlphaFold raises the tantalizing prospect that we may be able to incorporate relevant physical, chemical and biological constraints with neural networks to create new and better software tools and environments for advancing other areas of science. Lastly, with initiatives in the US and Europe we may be coming closer to realizing Jim Gray's dream of an open science world (Fig. 1).

Acknowledgement

This work was supported by EPSRC Grant EP/T001569/1 as part of the "AI for Science" theme of the Alan Turing Institute.

References

1. The National Oceanic and Atmospheric Administration OAA Data Discovery Portal. data.noaa.gov, last access: 15.10.2019
2. The big ocean button challenge. https://www.herox.com/bigoceanbutton/entries, last access: 15.10.2019
3. The Dryad Digital Repository. www.datadryad.org/, last access: 15.10.2019
4. The Zendo repository. https://zenodo.org/, last access: 15.10.2019
5. Alfred P Sloan foundation funding. http://blog.zenodo.org/2019/07/17/2019-07-17-dryad-partnership/, last access: 15.10.2019
6. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (2018) Multi-omics Factor Analysis – a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, https://doi.org/10.15252/msb.20178124, last access: 15.10.2019
7. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518:529–533, https://www.nature.com/articles/nature14236, last access: 15.10.2019
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vision 115:211–252, https://link.springer.com/article/10.1007/s11263-015-0816-y
9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118, https://www.nature.com/articles/nature21056
10. Fogel AL, Kvedar JC (2018) Artificial intelligence powers digital medicine. npj Digital Medicine, https://doi.org/10.1038/s41746-017-0012-2, last access: 15.10.2019
11. Hinton G (2018) Deep learning – a technology with the potential to transform health care. JAMA 320(11):1101–1102, https://doi.org/10.1001/jama.2018.11100, last access: 15.10.2019
12. Foster I, Gannon DB (2017) Cloud Computing for Science and Engineering. MIT Press, Cambridge London, https://mitpress.mit.edu/books/cloud-computing-science-and-engineering, last access: 15.10.2019
13. The National Infrastructure Commission. https://www.nic.org.uk, last access: 10/2019
14. The National Infrastructure Commission Data As Infrastructure Report. https://www.nic.org.uk/wp-content/uploads/Data-As-Infrastructure.pdf, last access: 15.10.2019

15. The National Infrastructure Commission Data for the Public Good Report. https://www.nic.org.uk/wp-content/uploads/Data-for-the-Public-Good-NIC-Report.pdf, last access: 15.10.2019
16. DAFNI: Data & Analytics Facility for National Infrastructure. https://www.dafni.ac.uk/, last access: 15.10.2019
17. Atkinson M, Gesing S, Montagnat J, Taylor I (2017) Scientific workflows: past, present and future. Future Gener Comp Sy 75:216–227
18. The Common Workflow Language. https://www.commonwl.org/, last access: 15.10.2019
19. The ELIXIR project. https://elixir-europe.org/, last access: 15.10.2019
20. The Oxford Research Archive. https://www.bodleian.ox.ac.uk/bdlss/digital-services/data-archiving, last access: 15.10.2019
21. Research Data at Springer Nature. https://researchdata.springernature.com/, last access: 15.10.2019
22. Nature Scientific data. https://www.nature.com/sdata/, last access: 15.10.2019
23. OECD Open Science. http://www.oecd.org/sti/inno/open-science.htm, last access: 15.10.2019
24. Fostering Open Science Project. https://www.fosteropenscience.eu/, last access: 15.10.2019
25. Schema.org. https://schema.org/, last access: 15.10.2019
26. Bioschemas. https://bioschemas.org/, last access: 15.10.2019
27. JSON for Linking Data. https://json-ld.org/, last access: 15.10.2019
28. Turing award winners. https://amturing.acm.org/, last access: 15.10.2019
29. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09, http://image-net.org/papers/imagenet_cvpr09.pdf, last access: 15.10.2019
30. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90, https://doi.org/10.1145/3065386, last access: 15.10.2019
31. Microsoft Research 2015 win. https://blogs.microsoft.com/ai/microsoft-researchers-win-imagenet-computer-vision-challenge/, last access: 15.10.2019
32. AlphaFold: Using AI for scientific discovery: https://deepmind.com/blog/alphafold/, last access: 15.10.2019
33. Evans R, Jumper J, Kirkpatrick J, Sifre L, Green TFG, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Jones DT, Silver D, Kavukcuoglu K, Hassabis D, Senior AW (2018) De novo structure prediction with deep-learning based scoring. In: Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts), 1–4 December 2018. Retrieved from: https://kstatic.googleusercontent.com/files/b4d715e8f8b6514cbfdc28a9ad83e14b6a8f86c34ea3b3cc844af8e76767d21ac3df5b0a9177d5e3f6a40b74caf7281a386af0fab8ca62f687599abaf8c8810f, last access: 15.10.2019
34. Geoffrey Hinton as quoted by Jeremy Howard in 2014. https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn, last access: 15.10.2019
35. OSTP Memorandum. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, last access: 15.10.2019
36. The Global Research Council statement on Open Access. https://www.globalresearchcouncil.org/fileadmin/documents/GRC_Publications/grc_action_plan_open_access_FINAL.pdf, last access: 15.10.2019
37. G8 Open Data Charter. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf, last access: 15.10.2019
38. Plan–S. https://www.coalition-s.org/, last access: 15.10.2019
39. 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 C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, 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. Sci Data 3:160018, https://doi.org/10.1038/sdata.2016.18, last access: 15.10.2019
40. Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M, the FAIRsharing Community (2019) FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol 7:358–367, https://doi.org/10.1038/s41587-019-0080-8, last access: 15.10.2019
41. FAIRsharing standards and policy site. https://fairsharing.org/, last access: 15.10.2019