IFLA WLIC 2019 Athens

Submitted on: 03.07.2019

Now or Never: Innovative Tools and Services for Scientists

Margret Plank
Technische Informationsbibliothek (TIB) – German National Library of Science and Technology, Hannover, Germany
Email address: margret.plank@tib.eu

Co-Authors:

Bastian Drees
Christian Hauschke
Angelina Kraft
Katrin Leinweber
Technische Informationsbibliothek (TIB) – German National Library of Science and Technology, Hannover, Germany

Abstract

In 2017 TIB carried out a survey in order to obtain a more accurate and multi-faceted picture of their target groups and their information and publication behaviour. The focus of the survey was on the use of non-textual scientific materials such as audio-visual media and research data. In this article, we present the implications of the survey results for the TIB. The survey revealed a high potential for services for non-textual materials like scientific videos, research data and research software. We introduce services like the TIB AV-Portal, the Leibniz Data Manager and Software Carpentries Workshops as well as services for interlinking of PIDs. In addition we present our VIVO System, which is a research profile system based on open source software in which scholarly output of an institution - including non-textual information - can be aggregated and made visible for the community.

Keywords

Non-textual Materials, Research Data, Scientific Videos, Scientific Software, PID Services, Infrastructures, AV-Portal, Software Carpentries, VIVO

Introduction

Today, scientific libraries function as modern information infrastructure facilities that claim to accompany researchers throughout the entire scientific process with innovative services. In order to meet this demand, libraries must continuously analyse the changing needs of the scientific communities, observe trends, implement technical innovations and develop target-group-oriented services. Moreover, the expectations of the users have to be regularly assessed in order to satisfy their requirements.

In 2017 the German National Library of Science and Technology (TIB) carried out a survey on the information acquisition and publication behaviour of scientists with a focus on non-textual materials.[1] The 1,400 responses from the study improve our understanding of the needs and requirements of our user groups, and contribute to a better adaptation of future developments to user demands. The results of the study confirmed some assumptions: Research data are a central part of scientific output; Open Access can contribute to increased visibility, but requires a comparable reputation of the publication organs; in addition to classical publications, other artefacts such as software, knowledge graphs, 3D models, videos and of course data are becoming increasingly important. Against this backdrop, we are constantly developing innovative tools and services, which support the scientists throughout the research life cycle. These include, for example, the

- AV-Portal for sharing scientific videos,
- the interdisciplinary data management system Leibniz Data for the processing, publishing and visualizing of data workflows,
- new DataCite services for interlinking PIDs such as DOIs and ORCiDs,
- trainings based on the internationally acclaimed Software, Data and Library Carpentries and
- the community-based research profile system VIVO

In this paper we will introduce exemplary TIB tools and services for non-textual materials, which were developed in close cooperation with the scientific communities according to a user-centred design approach. Additionally we will show how the interconnection of these services are maximising the benefits of our user groups. This applies in particular to our VIVO system, an open source research-focused discovery tool that supports recording, editing, searching, browsing, analysing, and visualizing scholarly activity and enables collaboration among scholars of all disciplines.

New Publication Formats

Our aim is to make any research output, beyond classical text publications, freely accessible and reusable. These are, on the one hand, the data on which the research results are based and the software used to generate or process them, and on the other hand, visualisations such as graphics, 3D models, animations, simulations or videos that enable understanding or interpret data.

For scientific libraries, many of these fields of action, such as research data management, are by no means new but have long been standard for most infrastructure facilities. Nevertheless, there is a great need for improvement here, for example with regard to the reusability as well as machine-readability and human readability of research data. At the same time, comparable activities in the field of scientific

[1] https://doi.org/10.22000/54

software, i.e. the management of research software, have hardly been established in libraries to date. In addition, media-specific solutions are required for other non-textual materials, such as audiovisual media or 3D objects (Strobel 2014), which are very rarely offered by libraries.

Institutions which provide infrastructures have the central task of "supporting man and machine in finding, accessing, integrating and analysing scientific data". The FAIR Data Principles, which were published by the FORCE11 Group in 2016 (Wilkinson 2016), can be understood as a guideline. In these principles, FAIR stands for Findable, Accessible, Interoperable and Reusable. Even though the FAIR principles were inspired by Open Science, these two movements are explicitly not to be equated (Mons 2017). However, the FAIR principles require "clarity and transparency about the conditions for access and re-use" of the data, even if they are not "open".

The role of libraries must on the one hand be to provide infrastructures which are based on the FAIR principles and try to continuously increase their 'FAIRness'. On the other hand, the library plays the role of the link between data producers* and (re-)users and can support both with its expertise, especially in the areas of metadata and licences. In the following chapter we briefly introduce individual TIB Infrastructures and services, which follow those principles.

TIB Infrastructures and Services

AV Portal

The TIB AV-Portal[2] is a web-based repository platform for quality-tested scientific videos, such as conference recordings, computer visualizations, simulations or video abstracts (Plank 2016). The portal was developed jointly by TIB and the Hasso Plattner Institute and is operated by TIB since 2014. Currently it offers access to approximately 20,000 videos, which are predominantly provided under Open Access licenses. The AV-Portal combines automated metadata generation based on automated video analysis (such as scene, speech, text and image recognition) as well as semantic search and cross-lingual retrieval (German/English). Based on these technologies, content can be localized within the videos using semantic and explorative search functions. The assignment of Digital Object Identifiers (DOI) and Media Fragment Identifiers (MFID) ensures a stable and precise citation capability. The metadata of the scientific videos from the AV-Portal is published as linked open data[3].

[2] https://av.tib.eu

[3] https://av.tib.eu/opendata

Figure 1: TIB AV-Portal (https://av.tib.eu/)

In order to operate a successful open video repository own research activities e.g. in the field of machine learning for video analysis are being transferred to the AV-Portal. At the same time we keep on engaging with the needs of our users e.g. by web analytics, user interviews or usability tests. In this way, possible weak points of the repository and potential for further development can be determined.

The constant contact with the user also revealed the need of the scientists for skills training regarding the production of high-quality videos as well as legal questions. We address these needs in a variety of workshops for the production of (especially) video abstracts and ePublishing. We also offer a conference recording service for conference organizers, where we provide a full service from legal advice through production to publication[4].

Leibniz Data Manager

Reproducibility ensures the validation of scientific findings with minimal effort. However, scientific digital artefacts, e.g., data and computational methods, need to be findable and accessible in order to be used. At TIB, we currently develop, test and improve a tool which significantly increases the findability and exploitation of research data and other scientific artefacts, the Leibniz Data Manager (LDM).

The LDM is a CKAN-based software distribution that allows for a method which is called ‘deep indexing’ of metadata and data across digital repositories, using the existing semantic tools like DCAT to map metadata standards to semantic vocabularies. This enables researchers to search and screen for data sets and other scientific artefacts across multiple digital repositories based on their metadata and explore their relevance for their own research. At present, the ecosystem of scientific data repositories consists of a large variety of available categories and types: Discipline specific repositories, interdisciplinary repositories, institutional repositories, and mixtures thereof. With this heterogeneity comes large variation in terms of data and metadata standards, APIs, file formats, licence information, archival- and publication guidelines, terms of re-use, and others. This is also the reason why a search across multiple repositories is considered a time consuming task to be carried out by researchers who want to re-use data, but are unsure where to look for it. With the LDM, we target interoperability

[4] https://wiki.tib.eu/confluence/display/conferencerec/TIB+Conference+Recording+Service

challenges that are shared among research institutes, infrastructures, universities and companies and offer a simple, small-scale and open software distribution which can connect digital repositories in a way such that data sets and other scientific artefacts will stay in their respective repositories, with the LDM providing an integrated view of the data sets archived by these repositories. As such, the LDM can aid in the transition from a publication- or article-based to an information-based (linked-data) research workflow.

Services for Interlinking of PIDs

DOIs[5] and ORCID IDs[6] are both important persistent identifiers in everyday research. While the DOI assignment makes an object, e.g. a research dataset, a video or a publication, permanently citable, linkable and searchable, an ORCID ID serves as an identifier for individuals to use with their name as they engage in research, scholarship, and innovation activities. By assigning DOI names and hosting the DataCite e.V.[7] business office the TIB ensures quality control, permanent availability and referencing as well as the verification and access of research data via the TIB portal. In addition, TIB together with the ORCID DE project jointly established and maintains the ORCID Germany Consortium. In order to register DOI names in cooperation with the TIB, an institution (institute, university, data centre) applies to the TIB as a so-called data center. Using this point of contact, TIB provides its (potential) data centers with information on new DataCite services, such as the auto-update service between DataCite and ORCID. This service enables DataCite (providing the permission of the researcher who owns the ORCID profile) to automatically push metadata to an ORCID record when an ORCID identifier is found in newly registered data entries which have DOI names. For the researcher, this means that the ORCID profile is automatically updated e.g. as soon as a new research article or data set is published within a journal or a data repository. This auto-update service between DataCite and ORCID thus demonstrates an interlinking both identifier systems that leads to an increased machine-readability, findability and accessibility of academic content and aids to the researcher's reputation.

Services Based on the Software, Data and Library Carpentries

"The Carpentries" are a volunteer organisation that introduce researchers (Wilson 2016), librarians and other knowledge workers to foundational computing and data analysis skills. They aim is to "empower these communities to use software and data in their own work and train others"[8]. Using collaboratively developed OERs, the Carpentries specifically teach the Unix shell (Bash) and programming languages such as R or Python to automate repetitive tasks, version control with Git (Hub/Lab) to collaborate on digital projects, the Structured Query Language (SQL) to store and access scientific data in useful ways, and other related skills in usually 2-days long workshops. The overall goal is to significantly increase the productivity of scientists and knowledge workers and spread digital, data and software literacy skills.

TIB aims to combine the Carpentries approach with theory and practice of the FAIR principles[9], with particular regards to the software "side of the coin" as expressed by our hosting of the "Software and

[5] https://www.doi.org/
[6] https://orcid.org/
[7] https://datacite.org/
[8] LibraryCarpentry.org
[9] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Services for Science" (S3) conference[10] in May 2017. Continuing this theme, we hosted a "FAIR Data & Software" workshop[11] in July 2018 which trained junior scientists in implementing the FAIR principles for research data and software management and development. The workshop focussed on applying respectively appropriate good practices in preparing, publishing and archiving both these important scientific objects with respect to their similarities and differences.

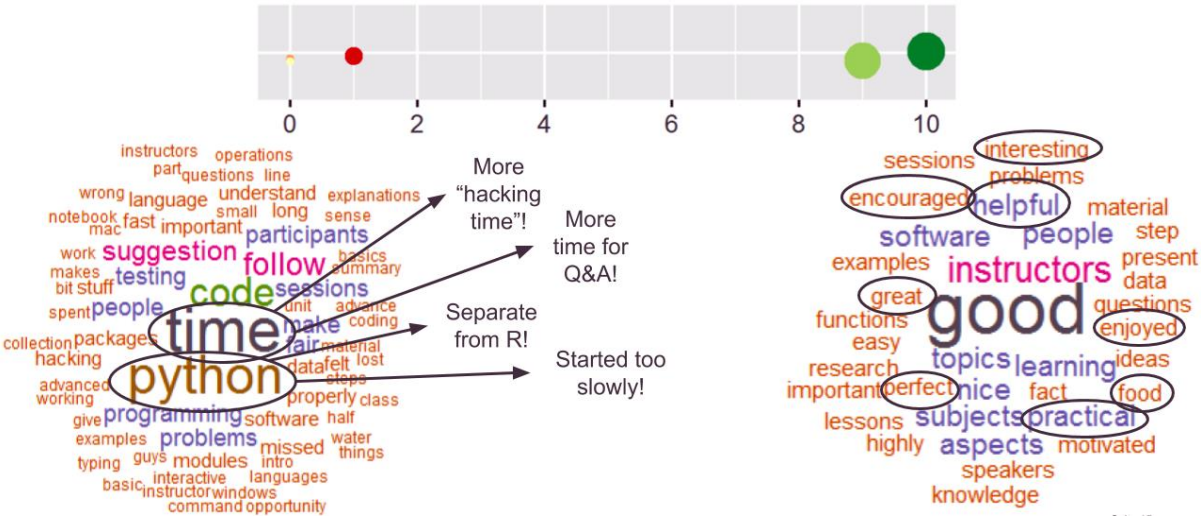

Figure 2: Summary of quantitative and qualitative feedback of TIB's "FAIR Data and Software" workshop; Numeric scale & dot size: number of replies; Color scale: overall rating (green = best); Wordclouds summarise written feedback (left = criticism; right = praise; highlights encircled)

The workshop was evaluated positively overall and let us conclude the following:

- Our own planning was improved by using GitHub Pages[12] to build and host the workshop website and by primarily using an issue tracker internally rather than email.
- 65% of the participants "would pay a small fee" according to our post-workshop survey. 20% said they wouldn't and 15% didn't answer. This inspires confidence that cost-covering fees are a viable alternative to seeking other funding sources.
- We prepared more own training materials than we reviewed existing ones. This likely biased our lesson style towards presenting more of what we thought important while discussing fewer of the data or software examples by the participants and providing too little time for 1-to-1 consulting for their own projects during the workshop. In future iterations, we plan to reduce the amount of custom material and increase the time spent on discussions and consulting.

To implement some of these conclusions in between especially prepared workshop events, we established a weekly "FAIR Study Group" to provide a permanent, peer-to-peer forum for these topics aimed at local students and researchers. Also, our recent workshop will focus on implementing the

[10] 2nd Conference on Non-Textual Information, see events.TIB.eu/nontextualinformation2017
[11] See events.TIB.eu/fair-data-software/2018 with recordings at av.TIB.eu/series/530
[12] See pages.GitHub.com plus template: Greg Wilson (ed). "Software Carpentry: Workshop Template." Version 2016.06, June 2016, GitHub.com/carpentries/workshop-template, 10.5281/zenodo.58156

FAIR principles using only the R programming language and the Git version control system as examples.

In the future, our engagement with the Carpentries will continue to include workshops, but expand beyond the natural sciences and library personnel to include teacher candidates and pedagogy students. A project is being planned, which aims at equipping these key target groups in the education sector with the above-mentioned foundational computing skills in combination with facilitating their training as Carpentries instructors.

VIVO

VIVO is an Open Source software to represent scholarly information. It is based on linked data, and persistent identifiers play a major rule in the reuse of data from other systems. VIVO can be used for a variety of use cases, for example as a platform to gather research information about an institution for research analytics and standardized reporting (Walther et al 2019), or to display the expertise of an institution and its staff via research profiles.

The TIB is involved in a variety of activities related to research information. TIB is running a publicly visible research profile system in which our scholarly active employees can create profiles in which their scholarly output can be aggregated and showcased. The aim is to portray the expertise of the TIB staff and thus of the TIB itself in a profile based on the research output of a person. In this way, not only will increased visibility of TIB activities be ensured, but the VIVO profiles will also be contact points for networking with people interested in these activities.

To integrate VIVO seamlessly with services like the AV-Portal, we have adapted a mechanism for importing publication data via identifiers like DOI or Pubmed ID originally developed for OpenVIVO (Ilik et al. 2018). This enables our researchers to claim basically any work with an assigned and registered DOI to be imported into our VIVO system. The described mechanism allows our researchers to import all kind of works with an assigned DOI, be it research data, software, video or - of course - conventional text-based publications like articles, chapters, or books.

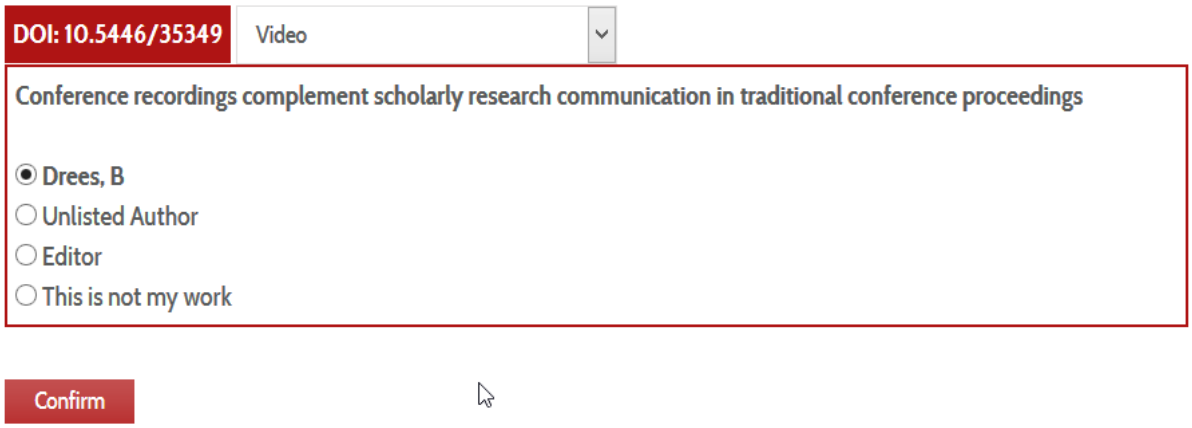

Figure 3: Claiming authorship for a video via the DOI import mechanism

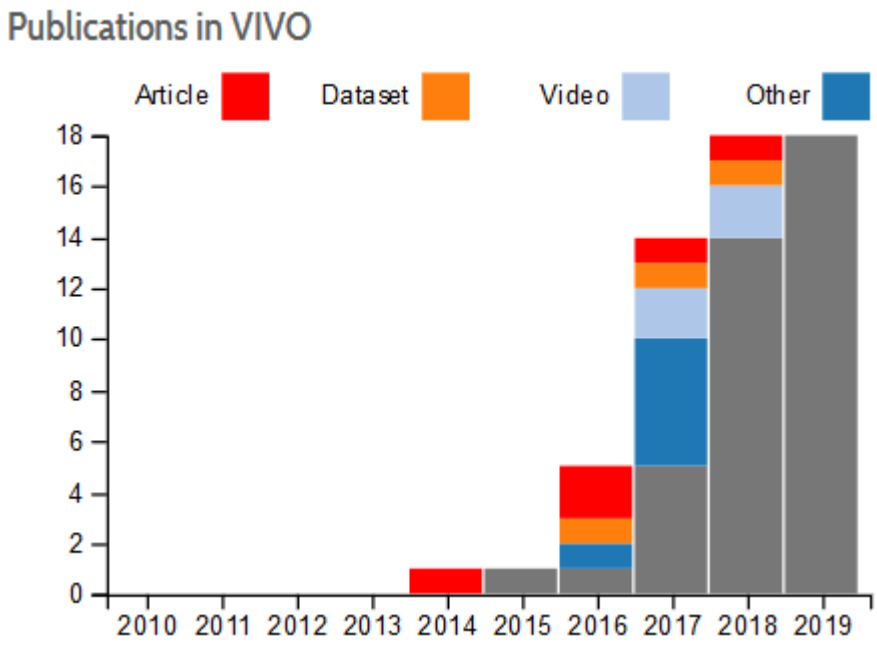

Figure 4: Articles, datasets and videos as publication types in the publication chart of a VIVO profile

Due to the very extensive and powerful VIVO ontology, it is possible to map a wide range of scholarly activities. In addition to the publication activities described here, this also includes participation in events such as the Carpentries described above.

Our objective is to cover the whole ground of research information in an open source system based on open data. To this end, we are working together with the VIVO community to further develop VIVO for ourselves and others and to help third parties use VIVO.[13]

Conclusion and Future Work

According to the study mentioned at the beginning of this article, different scientific status (with higher status and age here often positively correlated) often means different media behavior of scientists. In principle, it can be observed that both the frequency of the creation and use of non-textual materials as well as the described consulting needs are highest among PHD students. The higher the scientific status, the lower the use and need of these object types. These needs also vary greatly depending on the scientific discipline. For example, the handling of Open Access, Research data and repositories in the subjects of mathematics and science are more widespread and associated with fewer uncertainties than in many technical and engineering disciplines. It is not clear from the results whether the derived need for advice arises from the fact that the existing offers, e.g. in scientific libraries, are not sufficiently known by researchers and the marketing and sales activities have to be optimised accordingly, or whether the existing offers are better adapted to the different emerging needs of the target groups (depending on the state of scientific knowledge). In any case, we conclude from this that it offers great potential for scientific libraries, in particular, to collaborate with scientific institutions and to cooperate

[13] More information about our services and activities in the field of research information: http://tib.eu/researchinformation

more closely with communities and develop target group-specific formats and services. The TIB uses this potential and develops user-centred services for scientific videos, research data and software.

Strengthening Open Science means to give researchers incentives to do Open Science. This includes the opportunity to display different kinds of scientific activities to their respective communities, for example via research profiles. The traditional means to do this are not sufficient, too much emphasis is lying on conventional and highly criticized metrics like the Journal Impact Factor, or the H-Index. Future works includes e.g. the project ROSI (Reference implementation for Open Scientometric Indicators) where we are dealing with this topic. The main objective of the ROSI project is to develop a researcher friendly visualization of several dimensions of their research output which include the impact on scientific communities and society as well as openness.

References

Einbock, Joanna; Dreyer, Britta; Heller, Lambert; Kraft, Angelina; Niemeyer, Sandra; Plank, Margret; Schrenk, Philip; Sens, Irina; Struß, Julia; Tullney, Marco (2017): Informationsbeschaffungs- und Publikationsverhalten von Wissenschaftlerinnen und Wissenschaftlern der natur- und ingenieurwissenschaftlichen Fächern: Auswertung einer Umfrage mit Schwerpunkt auf nicht-textuellen Materialien. Hg. v. Technische Informationsbibliothek (TIB). Hannover. http://tib.fiz-karlsruhe.de/ir/item/escidoc:9102/components/component/escidoc:91023/conten

Hauschke et al. (2018): Reference implementation for open scientometric indicators (ROSI) https://doi.org/10.3897/rio.4.e31656

Ilik, V. et al. (2018): OpenVIVO: Transparency in Scholarship. https://doi.org/10.3389/frma.2017.00012

Mons, B., Neylon, C., Velterop, J., Dumontier, M., Da Silva Santos, L. and Wilkinson, M. (2017) 'Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud', Information Services and Use. 37 (1): pp. 49-56. doi: 10.3233/ISU-170824

Plank, M. (2016): TIB AV-Portal: A reliable infrastructure for scientific audiovisual media. In: Hana Vyčítalová and Michaela Charvátová (eds.): Conference on Grey Literature and Repositories Proceedings. 9th Conference on Grey Literature and Repositories Prague, 19 October 2016. Prague, National Library of Technology, p. 17-23. http://nrgl.techlib.cz/conference/conference-proceedings/t

Strobel, S. und Plank, M. (2014) 'Semantische Suche nach wissenschaftlichen Videos: Automatische Verschlagwortung durch Named Entity Recognition.' Zeitschrift für Bibliothekswesen und Bibliographie. 61. 255 - 259. http://dx.doi.org/10.3196/18642950146145154

Walther, T. et al. (2019): The Research Core Dataset (KDSF) in the Linked Data context. https://doi.org/10.1016/j.procs.2019.01.074

Wilkinson, M.D., et al. (2016) The fair guiding principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, G. (2016): Software Carpentry: lessons learned. doi.org/10.12688/f1000research.3-62.v2