



# Recommendations for an Open Science approach to welding process research data

Cagtay Fabry<sup>1</sup> · Andreas Pittner<sup>1</sup> · Volker Hirthammer<sup>1</sup> · Michael Rethmeier<sup>1,2,3</sup>

Received: 20 April 2021 / Accepted: 5 June 2021  
© The Author(s) 2021

## Abstract

The increasing adoption of Open Science principles has been a prevalent topic in the welding science community over the last years. Providing access to welding knowledge in the form of complex and complete datasets in addition to peer-reviewed publications can be identified as an important step to promote knowledge exchange and cooperation. There exist previous efforts on building data models specifically for fusion welding applications; however, a common agreed upon implementation that is used by the community is still lacking. One proven approach in other domains has been the use of an openly accessible and agreed upon file and data format used for archiving and sharing domain knowledge in the form of experimental data. Going into a similar direction, the welding community faces particular practical, technical, and also ideological challenges that are discussed in this paper. Collaboratively building upon previous work with modern tools and platforms, the authors motivate, propose, and outline the use of a common file format specifically tailored to the needs of the welding research community as a complement to other already established Open Science practices. Successfully establishing a culture of openly accessible research data has the potential to significantly stimulate progress in welding research.

**Keywords** Welding · Open Science · Research data management · Data exchange · FAIR · File format · Open Source · WelDX

## 1 Introduction

### 1.1 State of research data

In recent years, the shift towards a research landscape increasingly shaped by digitalization and Open Science principles is in full effect. Rather than a single principle, Open Science as a whole is commonly understood as a governing taxonomy of multiple related ideas, guidelines, and concepts such as Open Access, Open Reproducible Research, and Open Data. While each focus on different aspects of the scientific process, all principles combined aim to provide more access to scientific research practices,

data, and publications with the goal of ultimately leading to higher quality research results.

In some fields, this shift has been motivated and accelerated by increasing difficulties when it comes to reproducing published scientific findings [1, 2]. While independently reproducing previous experiments and results should certainly be emphasized in any field, the prior in-depth understanding of the existing data is a fundamental step and prerequisite in complex research applications such as welding. Without a thorough description of all relevant acting effects and boundary conditions, attempts at an accurate reproduction of previous experimental results will be limited. Overcoming these limitations of course requires access to the underlying data of previous work. As Wilkinson et al. point out [3], data should be made “machine actionable” as much as possible to facilitate the reuse of data and good scientific practices. One major building block towards Open Science—specifically Open Data—in this regard is the practice of the FAIR Guiding Principles (Findable, Accessible, Interoperable, Reusable) established by Wilkinson et al. [3]. In welding and related fields, applying the FAIR Guiding Principles to scholarly data is still an ongoing and challenging issue [4, 5]. The

✉ Cagtay Fabry  
cagtay.fabry@bam.de

<sup>1</sup> Bundesanstalt für Materialforschung und –prüfung (BAM),  
12205 Berlin, Germany

<sup>2</sup> Institute of Machine Tools and Factory Management,  
Technical University of Berlin, 10623 Berlin, Germany

<sup>3</sup> Fraunhofer Institute of Production Systems and Design  
Technology (IPK), 10587 Berlin, Germany

fundamental work to provide the necessary infrastructure to make and keep data FAIR is an ongoing effort by the broader scientific community. When it comes to making the data reusable for welding sciences however, the authors think that a crucial missing element that has to come from inside the community is the answer on how to represent, interchange, and archive experimental welding datasets.

The advent of large, high-quality, and openly accessible datasets has been one of the reasons machine learning and data analysis have been on the rise. In many fields curated reference datasets exist and are used by global research communities to develop and compare different solutions. Some of these datasets are highly complex—created and curated with extensive manual efforts. The Open Images datasets such as described in [6] contain millions of richly annotated images that are verified by humans and used for classifications. Another field with recent influx in scientific and industrial attention is the application of machine learning, image classification, and pattern recognition in the context of developments of self-driving cars. Examples are the extensive datasets such as the BDD100K annotated driving videos described in [7] or the detailed combined radar, lidar, and camera datasets put forth in [8]. Notably, all datasets are available under permissive licenses for scholarly use. The apt use and citation of these datasets in their scientific communities are a good example of recognition of the work put into creating, describing, and making the data available and providing incentives for the authors to publish further datasets.

Welding has long been an innovative field with early data-driven research and applications of machine learning, artificial intelligence, and similar techniques due to its inherent complexity [9–11]. However, it has proven difficult to keep up with current trends in digitalization and research data management due to some prevailing challenges in nature and conduct of welding science.

## 1.2 Situation in welding science

The general publication process in welding sciences with regard to scientific articles is well established and continues to transition to a more Open Access focused model in line with changes in related fields. The change towards Open Access and related practices such as Open Peer Review, while appreciated by the welding science community, is mostly driven and handled by external factors and institutions.

In an increasingly data-driven research landscape, welding research is no exception to current trends such as data fusion and machine learning applications [3, 11]. However, in contrast to other scientific fields with an increasing focus on data-driven research, there exist no publicly available welding research datasets that enable

independent evaluation and advancements of the mentioned methods. Due to the expensive and time-consuming experimental efforts needed to create high-quality datasets in welding-related fields, these datasets—especially the raw data—often are not made publicly available and remain as institutional knowledge. Consequently, possibilities to uncover or validate new findings based on aggregation of research data from multiple sources are limited.

One part of adopting Open Science besides the established peer review publication process that does require significant direct input—and maybe in part a change in ideology from the researchers and institutions themselves—is applying the FAIR Guiding Principles to welding research data. Like other fields, the welding community can use the underlying emerging infrastructure and platforms on a national or international level—like the German National Research Data Infrastructure for engineering (NFDI4ing) [12] or the European Open Science Cloud (EOSC) [13] initiatives respectively—as a general basis for applying their own ontologies<sup>1</sup> and metadata schemas [4]. However, even with the tools to make welding research data findable and accessible, one core challenge remaining is ensuring the comprehensibility and machine (re-) usability of complex welding research datasets. While it is technically possible to upload raw and processed welding research data as of now, some key practical, ideological, and technical challenges have historically hindered the adoption of Open Data practices specifically in welding research.

### 1.2.1 Practical challenges

Focussing on fusion welding, one of the main and most fundamental practical challenges is inherent to the way welding experiments are conducted at research institutes. Modern fusion welding laboratories are highly complex setups, often consisting of automated workpiece and torch manipulators, different welding power sources, and multiple specialized sensors and secondary monitoring equipment. Since most institutes also specialize in a particular field of expertise and according experimental setups, it is probably safe to say that no two identical experimental setups exist at different institutes. The diversification in arc and laser beam welding seems to be especially high since a plethora of welding equipment is available and custom research setups are widespread and usually easier to realize compared to other welding processes such as electron beam welding that are often more in line with industrial applications adapted for research purposes.

<sup>1</sup>In the context of information science, the term ontology is commonly used to describe “a formal, explicit specification of a shared conceptualization” [14, 15].

As a result of the complex experimental setups, researchers in welding science often face the challenge of having to work with many data sources and file formats specific to their field of application and equipment used. This further complicates the exchange of research data since some produced initial raw data files might not even be accessible without a specific commercial software or license. What specific file formats are used relies heavily on the setup of each institute. In a way, the situation seems comparable to the one described by Wells et al. back in 1979 [16] where different astronomical installations, while in principle focussing on the same research area, did use different setups, computing hardware architectures and software, and subsequently different internal file formats for describing and storing similar observations. Notably, the conclusion of Wells et al. was not to unify the internally used file formats, which would reduce efficiency and cause significant expenses, but define a suitable interchange format with the specific intent to act as an agreed upon data transferring and interchange method. The resulting “Flexible Image Transport System (FITS)” format—while not without shortcomings of its own [17]—has since been used successfully for many purposes in astronomy as an interchange and archival format with great success for decades.

### 1.2.2 Ideological challenges

Besides the practical challenges of data sharing, another important factor to consider is creating incentives and motivate researchers to make their data available. This is especially true considering the required initial upfront investment of time and funding money for realizing the necessary surrounding conditions.

So far, there have been no use cases based on and building on widely available welding research data, general data exchange, or Open Access data publishing in welding sciences. Exchange of raw data between institutions is mostly limited to collaborative projects. Unfortunately, no known exchange formats have been published as a result of collaborative projects so far to the authors’ knowledge.

Up to now, no Open Data platforms for welding research exist. In the rather competitive publication environment of welding research, investing into Open Data research practices as an early adopter or even originator, the risk might seem greater and rewards lesser. This holds especially true since data publications historically are often ranked considerably lower than peer-reviewed articles concerning their impact and reputation.

The lack of incentives for openly sharing data is one of the key hurdles that have to be overcome by the welding community as a collective effort. The authors think that

establishing a culture of openly accessible research data of high quality has the potential to significantly stimulate progress in welding research and open up new areas of research and expertise as well as pave the way for new collaborations.

### 1.2.3 Technical challenges

The technical challenges of agreeing upon a suitable transportation or archival format for experimental welding data mostly stem from the inherent complexity and diversity of the welding process and experimental procedures themselves.

Many welding processes such as arc welding are the result of complex occurrences happening concurrently in multiple physical and metallurgical domains, as is evident from increasingly elaborate simulation efforts. In addition, the relevant effects cover a wide range of different time scales, ranging from micro-second monitoring of the welding process itself to temperature, metallurgical, and mechanical observations spanning minutes, hours, or even longer. This diversity has to be considered on a technical level regarding synchrony, precision, and resolution. Dealing with manufacturing and machine tolerances can greatly affect the welding process and resulting weldment. Hence, the representation of real-world measurements, raw data, and aggregated information should be supported and preferred over pure design descriptions for scientific purposes whenever possible. Data-focused description of the welding process is further complicated when considering complex workpiece geometries, varying boundary conditions and spatial as well as time-dependent relations between the welding process and the sensors providing measurements. The resulting complex spatiotemporal relations require the implementation of flexible and powerful data models that should be reflected in a file format. Besides static metadata, raw measurement data in the form of multi-dimensional arrays such as tabular data (scalar time series), video frames or geometric—respectively pointcloud—data can represent most information and should be supported. Due to the often considerable size of raw data for experimental welding data, binary storage with optional compression of data seems like a sensible solution and is common in existing formats. The key challenge seems to lie in reflecting the flexibility and diversity found in experimental setups in an appropriate technical form.

In the authors’ opinion, defining and using a common open source file format can greatly contribute to promoting Open Science practices and ideals in welding research. Due to the complexity of welding research data, a common data format must be based on and incorporate a comprehensive but powerful underlying data model.

### 1.3 Previous efforts on data models for fusion welding

Regarding welding, considerable effort has been put into structuring and describing welding applications and related fields such as testing of weldments. Some of these efforts are finalized in standards that are used in industrial production and a critical aspect of safety considerations for welded components. In addition, some publications describe suitable data models for a limited range of applications.

The publication by Rippey [18] is most notable for its extensive scope, covering data models for weldment and joint descriptions, arc welding process specifications, and welded products including destructive weld inspections. Rippey mentions XML as a suitable example file format for a possible implementation basis of the proposed data model. In essence, Rippey provides a data model that combines multiple standards into a complete and comprehensive, uniform description. While the presented data model is complete regarding its scope and tightly connected with the American Welding Societies' (AWS) nomenclature and standards, the lack of a concrete implementation also means there are no practical application examples provided. With regard to contents and applicability for research, the omission of handling time-dependent metadata or more complex data models for concepts of measurements, coordinate transformations, and their relation to welding data is limiting.

In addition to defining the data model, Rippey adds—among other recommendations—the following valuable “tasks for data modelers” for proceeding towards an actual database or data format implementation [18]:

1. Specification of required, optional, or forbidden relations between data models depending on different applications (such as GMAW) or use cases
2. The implementation be agnostic with regard to customer-defined units, i.e., handle both SI and U.S. customary units
3. Modularize the provided welding schema definitions into smaller, distinct parts

Another effort was put forth by Kristiansen [19, 20]. In contrast to Rippey's AWS-based data model, Kristiansen emphasizes creating a generic information model for describing empirical data from the welding process from a researcher's perspective with considerable effort to represent a complex automated welding environment. One goal of the data model is to derive process information and experimental data in a form and level of detail suitable for machine learning applications. This is a similar and common challenge seen in many of today's machine learning applications to engineering fields. The introduction of so-called welding experiment samples as the smallest

examined domain making up the weld seam and forming the welding experiment allows the consideration of discrete points of time and space along the workpiece as a notable extension of the model from Rippey. In addition, a description of welding tool-frames, their spatiotemporal relation to the workpiece, and weld seam orientation as well as their mathematical representations are presented. The model lends itself adequately to produce and represent the gathered empirical data in a tabular manner that can be processed further and be used for machine learning tasks as demonstrated in [20]. The condensed data records are presented in the Appendix; however, the underlying raw data is seemingly not publicly accessible. The steps taken are described at length and illustrate the complexity of the task even for a fixed and well-defined single welding environment and procedure with single-layer weldments for two different groove types.

So far, two key elements that have been lacking and prevent a more widespread adaptation of these ideas and their practical application have been the following:

1. A practical and publicly available file format definition and implementation of the aforementioned data models. Collaborative platforms have emerged and established themselves over the last decade in the research context that greatly streamline the needed processes and efforts to maintain such a project. The implementations of previous work at different institutions would certainly provide a good starting point but are seemingly not publicly accessible.
2. Providing a user-friendly approach to the agreed upon data model and file format is essential. For this purpose, one or more suitable accompanying application programming interfaces (API) should be provided. Ideally, the API should simplify the use of the file format and lower the entry barrier considerably. Science nowadays being more data driven also means a more programming-focused approach to many problems, solutions, and tools. This should be reflected in the way a modern API for welding applications is designed.

## 2 Advantages of using a common file format

The use of a common file format for research data can potentially provide diverse advantages in different stages and aspects surrounding the research data life cycle as has been practically proven by applications in numerous fields such as crystallography [21, 22], climate research [23], ecological sciences [24], proteomics [25], earth sciences [26], and medical imaging [27]. Many of these advantages are inherent to the concepts of Open

Science practices but should be mapped out in detail for welding research.

One immediate and major advantage of a consolidated effort lies in providing a single place for the community to provide documentation and discussion. Even if effort is needed to establish an initial point of origin, subsequent work and extensions of the documentation should come more easily and visible. Building upon existing work in an iterative process rather than creating isolated solutions will be beneficial in the long term. As has been shown in astronomy [17, 28],<sup>2,3</sup> biophysics [29],<sup>4</sup> molecular dynamics [30],<sup>5</sup> materials science [31]<sup>6</sup>, and numerous other fields, existing Open Science and Open Source workflows are already well established and promote an open discussion culture suitable for research communities.

Providing an agreed upon approach will reduce the documentation effort needed for each individual researcher and publication. Currently, considerable time and effort need to be taken to lay out and explain in detail the experimental design and related measurements for each publication. This holds especially true for research of complex dynamic or automated welding experiments [11] and oftentimes requires a trade-off between completeness and conciseness. With a suitable and well-documented data model and format, the burden of documenting experimental procedures for peer-reviewed articles should become much more concise and precise at the same time. Thus, leaving more space for prolonged discussion of the researchers findings without sacrifices in clarity of experiments. On the contrary, new and universal tools can be applied to greatly enhance accessibility and understandability.

Ultimately, this effort should reduce documentation overhead for individual researchers, institutions, and the welding community as a whole and in addition ease collaborative effort between institutions. One obvious use case would be to facilitate cooperation between institutes that have extensive experimental capacities, and those that focus on numeric or analytic simulations of welding processes, by providing precise and robust experimental datasets for model validation. Wei et al. demonstrated the modeling of grain structure evolution for GTAW of aluminum in [32] where modeling results could be validated based on independent experimental data presented by Schempp et al. [33]. In turn, those calibrated models may be used again to better understand and design

further experimental work. A shared pool of welding knowledge should decrease overhead and redundancy in any case and lead to reproducible reference datasets and results.

One should also highlight the importance of providing documentation from inside the welding community as opposed to simply referring to best practices in other fields. Documentation provided and discussed by the welding community will be more accessible for members of that community compared to other more generic approaches. In addition to being a work of reference, the documentation could be a more lively and dynamic place for discussion between people of different backgrounds and knowledge or experience levels. In the future, a common file format might even be used for educational purposes early in the qualification process of new researchers.

On a more practical level, using a common file format and related data models are essential steps towards the goal of accumulating a vast and useful pool of historic welding data and knowledge. This could open up new avenues for research while in turn reducing experimental and research expenses and repetitions. Existing historical data could be used for validating new data and be considered during development of new methodical approaches. As pointed out by Rippey, the definition and application of suitable data models are necessary to preserve accessibility to welding data and—particularly—knowledge [18].

Unified abstract data models and file formats allow new tools to be developed quickly, and the research efforts to keep up with, participate, and benefit from advancements in other fields. Successful developments and tool ecosystems have emerged in multiple fields [29, 31, 34, 35]. In the authors' opinion, welding sciences could benefit greatly from participating in similar environments.

In the future, the widespread adoption of a common file format might also lead to faster adoption rates from research findings into industrial applications or standardization practices. Conversely, the continuing trend towards digitalization and machine-actionability of standards [36] provides an additional starting point for integration and can be reflected into data models for welding research data. Using a common file format as a bridge between novel complex research applications and digitized standards could help foster relations between research and standardization in both directions. The format or an adequate derivate could even be integrated into industrial welding equipment and productive applications and environments.

Besides fields of possible applications, the main advantage as seen by the authors is to make an important step towards Open Data and fostering Open Science practices in welding research. In turn, this will help increase the quality of research data, publications, and findings by increasing comparability throughout the field.

<sup>2</sup><https://github.com/astrophy>

<sup>3</sup><https://github.com/asdf-format>

<sup>4</sup><https://github.com/markovmodel/PyEMMA>

<sup>5</sup><https://github.com/econ-ark>

<sup>6</sup><https://github.com/pyiron>



## 3 Data format for welding research

### 3.1 Technical considerations

When thinking about implementations of a suitable data format, fortunately research data in welding science does not commonly push technical boundaries—such as file sizes or access speed requirements—compared to other fields. In consequence, a practicable approach could be based on existing and well-adopted tools and solutions found and used in other scientific communities. Unstructured or mostly flat data formats such as CSV, `parquet`<sup>7</sup>, or `zarr`<sup>8</sup> are often found and excel in connection with describing single sources of data or in the context of computing intensive tasks. However, they seem too inflexible to adequately describe the needed correlation of multiple data sources and representation of welding experiments. Focussing on an archival format, a hierarchical representation of data appears a fitting approach. Such representation is—among others—provided and used by multiple existing data formats such as the Hierarchical Data Format (HDF5) [37], the Network Common Data Format (`netCDF`) [38], or the Advanced Scientific Data Format (ASDF) [17].

### 3.2 Design considerations

The design of a common file format should take the emphasis of the research communities' core requirements and domain-specific needs into consideration. A proven model for scientific file formats in complex applications has been the combination of a suitable base file format with incorporation of domain-specific data models or schemas. Some notable examples of this approach in other domains are:

- The NeXus format [21]<sup>9</sup> used in neutron, X-ray, and muon science, which builds upon the widely used HDF5 file format with domain-specific data models and schema definitions defined in external XML files. Work on the specification of the NeXus format is governed by the “NeXus International Advisory Committee”<sup>10</sup> with members from various international research institutes. The code and software are available on GitHub<sup>11</sup>
- The `netCDF` format which is used in conjunction with the Climate and Forecast (CF) metadata conventions in earth sciences [23]. Coordination of the work has

<sup>7</sup><https://parquet.apache.org/>

<sup>8</sup><https://zarr.readthedocs.io/>

<sup>9</sup><https://www.nexusformat.org/>

<sup>10</sup><https://www.nexusformat.org/NIAC.html>

<sup>11</sup><https://github.com/nexusformat>

moved from the original authors to a community governance structure [39].

- The “Flexible Image Transport System” (FITS) or even more so its proposed modern successor ASDF [17] that follows a more integrated approach to combining a file format with structured metadata and schema descriptions.

Building upon these existing solutions, the primary challenge appears not to be saving or describing single elements like time series or video data. The challenge rather lies in bringing all welding-related data and the represented physical domains into context in a cohesive manner. One proven approach has been the separation of hierarchical and structural definitions in schema files to represent the underlying data models [17, 21]. The data file contents are then referenced and validated against the schema specifications. This approach seems suitable to quickly integrate present welding standards and ontologies into research work and applications. In their motivation towards designing the ASDF format, Greenfield et al. make a strong case for human-readable formats over purely binary representations of contents and hierarchies. This may bear even more importance when considering long-term archival use cases. Moreover, a readable, intuitive, and approachable representation of the file contents but especially the schema and data model descriptions is desirable for ease of use and accessibility throughout the entire welding community. The schema descriptions should be easy to discuss, create, extend, or modify for different use cases by community members. If possible, no extensive technical knowledge should be required for participation to invite collaboration and increase flexibility. A growing and evolving data model description used throughout the community might prove to represent the most important step since adaptation to different file formats could follow much more quickly. Following scientific best practices, adequate ways of describing data provenance on different levels such as physical measurement chains and data and signal processing or the version of different file iterations and changes should be provided.

### 3.3 Collaborative considerations

Perhaps even more important than the technical and design aspects of a possible file format are the implications and ideas around building and promoting the data sharing mindset in the welding research community. Collaborative efforts among researchers on platforms like GitHub<sup>12</sup> or GitLab<sup>13</sup> under open source compatible licenses have been

<sup>12</sup><https://github.com/>

<sup>13</sup><https://gitlab.com/>

an emerging trend for many years and build the foundation of many rapidly developing projects especially in areas concerning data science. This is accelerated further by the rise of suitable and openly available software like the Python, R, or Julia ecosystems.

In the same vein, a welding science data format should be openly accessible under an open source license to foster collaboration and allow creation of distinct variations. Well-established code and discussion-related workflows seem suitable for scientific purposes [29, 31, 34]. They can easily be adopted to reflect numerous structures and decision-making processes of the welding community and existing associated governing bodies to achieve consensus.

Documentation should first and foremost address researchers involved with and working with the file format adequately to promote adoption. Lowering the entry threshold by providing meaningful and accessible documentation is an important goal when it comes to bridging gaps and bringing together researchers from different fields. Ideally those could be provided in the form of hands on or interactive learning materials and tutorials where appropriate. It should exist in an environment that allows direct and flexible communication to build an inclusive and welcoming but productive community. In a way, a “living” format could provide an additional, faster, and more accessible vehicle when it comes to sharing not only research data but also associated findings and results in addition to existing peer-reviewed publications.

Based on the abovementioned challenges and ideas, the Bundesanstalt für Materialforschung und -prüfung (BAM) has initiated the “Welding Data Exchange Format (We1DX<sup>14</sup>)” as an initial effort to create an open source file format. The project is targeted at arc and laser beam welding research applications and strives to provide a modern file format based on the scientific Python ecosystem with support for custom quality standard definitions. Implementation details of the file format will be provided in future publications, whereby an initial reference dataset has been published and presented in [40].<sup>15</sup> The ongoing development is open for collaboration on GitHub.<sup>16</sup>

Aside from all aspects directly related to creating and maintaining a file format, the welding community as a whole will have to face the challenge of finding ways to recognize and appreciate the work that will undoubtedly be required to publish high-quality datasets to sustain a lasting and valuable effort. As the main welding-related scientific journals and institutions are an important factor in the research landscape, the assessment of the International Institute of Welding (IIW) to acknowledge the importance

of the topic and continue the ongoing collaborative efforts in the context of We1DX is appreciated by the authors.

## 4 Summary and outlook

Digitalization and applying the Open Science principles in welding sciences represent an ongoing effort that requires multiple key elements to make progress. The idea of a common file format for research purposes to store and archive experimental welding data can be one of many steps to bring the welding community closer to the implementation of these ideas.

Following approaches in other fields, the authors recommend working on establishing a common file format tailored to the specific needs and requirements of the welding research community. Agreeing on a format for exchange and archival of welding data and knowledge can provide multiple valuable advantages. Most of all, accessibility of welding research data can be increased for peer-reviewed or data-focused publications. Furthermore, necessary individual as well as institutional efforts can be reduced by providing a single reference for documentation, development, and discussion. In addition, use of a common file format for welding data could increase distinctness and comparability of experimental works. In the long term, one goal could be to create and build upon expansive datasets and archives including research data from many institutions that each focus on different use cases such as welding process properties, welding simulation, or material effects. These datasets could be the foundation for more data-driven research applications or even be used in modern welding science-related education.

To aptly represent the complex relations of fusion welding processes in scientific environments and with the required precision, a file format should be extensive enough to not only store data but welding knowledge in the form of data models representing the spatiotemporal relations of the complete welding process. To achieve this, combining, adapting, and extending the existing approaches towards data models against the background and requirements of modern software development and data science tools represent a promising approach.

From a technical perspective, many of the required building blocks such as suitable base formats, programming and data analysis frameworks, or visualization toolkits already exist either inside or outside the welding community in the form of suitable base file formats and data modeling approaches. Hierarchical formats in combination with external schema definitions representing the underlying data models remain a well-established concept. To provide access and usability for many members of the welding community, a modern and open API to access and work

<sup>14</sup><https://bam.de/weldx>

<sup>15</sup>[https://github.com/BAMWe1DX/IIW2021\\_joint\\_intermediate\\_CXII](https://github.com/BAMWe1DX/IIW2021_joint_intermediate_CXII)

<sup>16</sup><https://github.com/BAMWe1DX/weldx>

with the file format seems equally important. Providing all parts of the project on modern collaborative platforms under a permissive open source license allows widespread collaboration and ensures future extensibility.

Such a coherent effort from inside the welding community will also help in providing incentives to collaborate on, share, and publish more welding datasets. The existence and accessibility of such datasets can lead to novel insights and use cases spanning welding knowledge from multiple institutions. Consequently, this might create incentives to more frequently publish welding data if these efforts are adequately appreciated in the community. As such, striving towards the use of a common and open file format for welding data appears a worthwhile endeavor.

To effectively collaborate on data exchange inside the research community, appropriate workflows and consortia will have to be established. In an initial step, publishing new or previously internal project results and software—e.g., internal guidelines and file standards with a focus on welding research data—for collaboration should increase visibility for the topic and help consolidate existing efforts. This could additionally be an important step to support rapid exchange of ideas and concepts between researchers, ideally gathering input and bringing together people from different backgrounds in the welding community. In this context, BAM will continue to publish the WeLDX codebase and results on GitHub. Subsequently, larger, more mature, and extensive solutions should emerge. At that point, national or international committees can take over a more guiding role under chair of one or multiple contributing institutes. In the long run, and to promote integration of research data standards with the broader welding community, these efforts and committee works could also be governed by existing more broadly based bodies in due form, e.g., an IIW working group or similar.

**Funding** Open Access funding enabled and organized by Projekt DEAL. The authors received the support from the Federal Ministry of Education and Research of Germany in the framework of WeLDX under project number 16QK12.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright

holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Baker M (2016) Reproducibility crisis. *Nature* 533(26):353–66
- (2015) Open Science collaboration: estimating the reproducibility of psychological science. *Science* vol 349(6251). <https://doi.org/10.1126/science.aac4716>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE et al (2016) The fair guiding principles for scientific data management and stewardship. *Sci Data* 3(1):1–9
- Franke S, Paulet L, Schäfer J, O'Connell D, Becker MM (2020) Plasma-mds, a metadata schema for plasma science with examples from plasma technology. *Sci Data* 7(1):1–11
- Fabry C (2019) Digitalization and open science in welding research. <https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/49381>
- Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Mallocci M, Kolesnikov A et al (2018) The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. [arXiv:1811.00982](https://arxiv.org/abs/1811.00982)
- Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, Madhavan V, Darrell T (2020) Bdd100k: a diverse driving dataset for heterogeneous multitask learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 2636–2645
- Meyer M, Kusch G (2019) Automotive radar dataset for deep learning based 3d object detection. In: *2019 16th european radar conference (EuRAD)*, pp 129–132
- Vasudevan M, Muruganath M, Bhaduri A, Raj B, Rao KP (2004) Bayesian neural network analysis of ferrite number in stainless steel welds. *Sci Technol Weld Join* 9(2):109–120
- Ghanty P, Paul S, Roy A, Mukherjee DP, Pal NR, Vasudevan M, Kumar H, Bhaduri AK (2008) Fuzzy rule based approach for predicting weld bead geometry in gas tungsten arc welding. *Sci Technol Weld Join* 13(2):167–175. <https://doi.org/10.1179/174329308X271751>
- Fabry C, Pittner A, Rethmeier M (2018) Design of neural network arc sensor for gap width detection in automated narrow gap GMAW. *Weld World* 62(4):819–830. <https://doi.org/10.1007/s40194-018-0584-8>
- Schmitt RH, Anthofer V, Auer S, Başkaya S, Bischof C, Bronger T, Claus F, Cordes F, Demandt E, Eifert T, Flemisch B, Fuchs M, Fuhrmans M, Gerike R, Gerstner EM, Hanke V, Heine I, Huebner L, Iglezakis D, Jagusch G, Klinger A, Krafczyk M, Kraft A, Kuckertz P, Küsters U, Lachmayer R, Langenbach C, Mozgova I, Müller MS, Nestler B, Pelz P, Politze M, Preuß N, Przybylski-Freund MD, Rißler-Pipka N, Robinius M, Schachtner J, Schlenz H, Schwarz A, Schwibs J, Selzer M, Sens I, Stäcker T, Stemmer C, Stille W, Stolten D, Stotzka R, Streit A, Strötgen R, Wang WM (2020) Nfdi4ing - the national research data infrastructure for engineering sciences. *Zenodo*. <https://doi.org/10.5281/ZENODO.4015201>
- Ayris P, Berthou JY, Bruce R, Lindstaedt S, Monreale A, Mons B, Murayama Y, Södergård C, Tochtermann K, Wilkinson R (2016) *Realising the European open science cloud*. European Union, Luxembourg. <https://doi.org/10.2777/940154>
- Studer R, Benjamins V, Fensel D (1998) Knowledge engineering: principles and methods. *Data Knowl Eng* 25(1):161–197. [https://doi.org/10.1016/S0169-023X\(97\)00056-6](https://doi.org/10.1016/S0169-023X(97)00056-6)



15. Staab S, Studer R (2010) Handbook on ontologies. Springer Science & Business Media, New York
16. Wells DC, Greisen EW (1979) Fits—a flexible image transport system. In: Image processing in astronomy, p 445
17. Greenfield P, Droettboom M, Bray E (2015) Asdf: A new data format for astronomy. *Astron Comput* 12:240–251. <https://doi.org/10.1016/j.ascom.2015.06.004>
18. Rippey WG (2004) A welding data dictionary. US Department of Commerce National Institute of Standards and Technology
19. Kristiansen M, Madsen O (2005) Process-planning models for welding using Bayesian network. In: 7th international conference on trends in welding research. ASM International, pp 635–640
20. Kristiansen M (2007) Modelling of the welding process using Bayesian network and applying data collected from several sources. Ph.D. thesis, Department of Mechanical and Manufacturing Engineering Aalborg University
21. Könnecke M, Akeroyd FA, Bernstein HJ, Brewster AS, Campbell SI, Clausen B, Cottrell S, Hoffmann JU, Jemian PR, Männicke D, Osborn R, Peterson PF, Richter T, Suzuki J, Watts B, Wintersberger E, Wuttke J (2015) The NeXus data format. *J Appl Crystallogr* 48(1):301–305. <https://doi.org/10.1107/S1600576714027575>
22. Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (cif): a new standard archive file for crystallography. *Acta Crystallogr Sect A* 47(6):655–685. <https://doi.org/10.1107/S010876739101067X>
23. Eaton B, Gregory J, Drach B, Taylor K, Hankin S, Caron J, Signell R, Bentley P, Rappa G, Höck H et al (2003) Netcdf climate and forecast (cf) metadata conventions. <https://cfconventions.org/cf-conventions/cf-conventions.pdf>. Accessed 27 May 2021
24. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M., Giovanni R, Robertson T, Vieglais D (2012) Darwin core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7(1):e29715
25. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci* 85(8):2444–2448
26. Krischer L, Smith J, Lei W, Lefebvre M, Ruan Y, de Andrade ES, Podhorszki N, Bozdağ E, Tromp J (2016) An adaptable seismic data format. *Geophys Suppl Mon Not R Astron Soc* 207(2):1003–1011
27. Mildenerberger P, Eichelberg M, Martin E (2002) Introduction to the dicom standard. *Eur Radiol* 12(4):920–927
28. Price-Whelan AM, Sipőcz B, Günther H, Lim P, Crawford S, Conseil S, Shupe D, Craig M, Dencheva N, Ginsburg A et al (2018) The astropy project: building an open-science project and status of the v2.0 core package. *Astron J* 156(3):123
29. Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JH, Noé F (2015) Pyemma 2: a software package for estimation, validation, and analysis of markov models. *J Chem Theory Comput* 11(11):5525–5542. <https://doi.org/10.1021/acs.jctc.5b00743>
30. Carroll CD, Kaufman AM, Kazil JL, Palmer NM, White MN (2018) The econ-ark and hark: open source tools for computational economics. In: Proceedings of the 17th python in science conference, pp 25–30
31. Janssen J, Surendralal S, Lysogorskiy Y, Todorova M, Hickel T, Drautz R (2019) Neugebauer, J.: pyiron: an integrated development environment for computational materials science. *Comput Mater Sci* 163:24–36
32. Wei H, Elmer J, DebRoy T (2017) Three-dimensional modeling of grain structure evolution during welding of an aluminum alloy. *Acta Mater* 126:413–425. <https://doi.org/10.1016/j.actamat.2016.12.073>
33. Schempp P, Rethmeier M (2015) Understanding grain refinement in aluminium welding. *Weld World* 59(6):767–784
34. Robitaille TP, Tollerud EJ, Greenfield P, Droettboom M, Bray E, Aldcroft T, Davis M, Ginsburg A, Price-Whelan AM, Kerzendorf WE et al (2013) Astropy: a community python package for astronomy. *Astron Astrophys* 558:A33
35. Eynard-Bontemps G, Abernathey R, Hamman J, Ponte A, Rath W (2019) The pangeo big data ecosystem and its use at cnes. In: Big Data from Space (BiDS'19), pp 49–52. <https://doi.org/10.2760/848593>
36. Loibl A, Manoharan T, Nagarajah A (2020) Procedure for the transfer of standards into machine-actionability. *J Adv Mech Des Syst Manuf* 14(2):1–13. <https://doi.org/10.1299/jamdsm.2020jamdsm0022>
37. The HDF Group (Unknown Month 1997) Hierarchical data format, version 5. <https://www.hdfgroup.org/HDF5/>. Accessed 27 May 2021
38. Rew R, Davis G (1990) Netcdf: an interface for scientific data access. *IEEE Comput Graph Appl* 10(4):76–82
39. Lawrence B, Drach R, Eaton B, Gregory J, Hankin S, Lowry R, Rew R, Taylor K (2006) Maintaining and advancing the cf standard for earth system science community data. <http://cfconventions.org/Data/cf-documents/cf-governance/cf2.whitepaper.final.pdf>
40. Fabry C, Pittner A, Hirthammer V, Rethmeier M (2021) Weldx – progress report on the welding data exchange format. <https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/52661>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.