Preparing the past for the future

Curating a daylight simulation model of Hagia Sophia for modern data infrastructures

Andreas NOBACK, Technische Universität Darmstadt, Germany
Lars Oliver GROBE, Lucerne University of Applied Sciences and Arts, Switzerland

Abstract: Digital humanities and artificial intelligence applications rely on structured sets of data and metadata. Larger and more complex dataset as a product of growing computing power and widely available advanced tooling demand modern data-management within projects, while scientific transparency, reusability, and interoperability demand machine-readable publishing and linking of project data. The latter implies infrastructure for long-term storage and AI-assisted search. Emerging platforms as the National Research Data Infrastructure (NFDI) and Specialised Information Services (FID), both sponsored by the German government provide suitable repositories but require the curation of research data. This paper examines the interdependency of in-project data-management and publishing and localizes possible AI-applications in in the context of a non-ideal case-study – the heterogeneous dataset of a light-simulation model of Hagia Sophia. It proposes a separation of the project-data into five scopes – raw data collections, reconstruction and material models, simulation environment, simulation results and digital publications – that allow to develop transferable solutions and integration into emerging infrastructures for reuse, search, linking and publishing of data between projects. The paper concludes that AI-applications in this context provide more general, transferable solutions for search and spatial image organisation within the research infrastructure and very specific solutions within a project. A standardised organisation of research data and metadata has to fit these applications.

Keywords: *Data Management—FAIR Principles—Simulation Model—FID—NFDI—AI Application*

CHNT Reference: Noback, A. and Grobe, L. O. (2020). 'Preparing the past for the future: Curating a daylight simulation model of Hagia Sophia for modern data infrastructures', in Börner, W., Rohland, H., Kral-Börner, C. and Karner, L. (eds.) *Proceedings of the 25th International Conference on Cultural Heritage and New Technologies, held online, November 2020.* Heidelberg: Propylaeum. doi:10.11588/propylaeum.1045.c14476

Introduction

A detailed simulation model of Hagia Sophia (Fig. 1) emerged from more than 20 years of research in its daylighting (Hauck et al., 2013; Noback et al., 2020).[1] From the start this model was meant for distributed development and sharing (Grobe et al., 2020). It seems to fit the recently available infrastructures for data publishing, but efforts to publish its heterogeneous data set reveal interesting challenges that – in the first step – demand some conceptional work. These challenges seem to be typical for the current state of digitalisation. Its resulting data-rich research environments demand

[1] The research started as a project with the title "Die Hagia Sophia Justinians in Konstantinopel als Schauplatz weltlicher und geistlicher Inszenierung in der Spätantike", that was founded by the Deutsche Forschungsgemeinschaft (#5194526).

data management to guarantee standard conformity, transparency, long term storage, accessibility, intellectual property and security. Adequate data publishing infrastructure is a requirement for the application of computational agents including *artificial intelligence* that depend on structured and machine-readable data and meta-data. This is best summarised in the FAIR Principles (Wilkinson et al., 2016) that require that "all research objects should be *Findable, Accessible, Interoperable and Reusable* (FAIR) both for machines and for people". Internal workflows, separation of different types of data, annotation, and external referencing have to comply with this kind of research practice.

Fig. 1. A visualisation of Hagia Sophia's interior in the sixth century accounting for contrast and brightness based on daylight simulation. The simulation environment includes a reconstruction of the historic geometry and the optical properties of the surfaces. The environment is associated with a multitude of research data. © Authors.

Data publishing in the context of new research data management platforms

Multiple platforms provide opportunities for data publication, research data management, and community integration. They ask for contributions from the research community to their further development. The German government alone invests in two such major research infrastructure programs:

1. The National Research Data Infrastructure (NFDI), whose "aim is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally."[2] It provides its science-driven data services to specific *research communities* through the consortiums NFDI4culture (art history, architecture etc.), NFDI4objects (archaeology, in preparation) or NFDI4Ing (engineering).

[2] https://www.dfg.de/en/research_funding/programmes/nfdi/

2. Through the Specialised Information Services (FID), the DFG funds[3] libraries that support scientists in specific, similar fields, e.g. arthistoricum.net[4] (art history), Propylaeum[5] (classical and ancient studies), or FID BAUdigital (building science).[6]

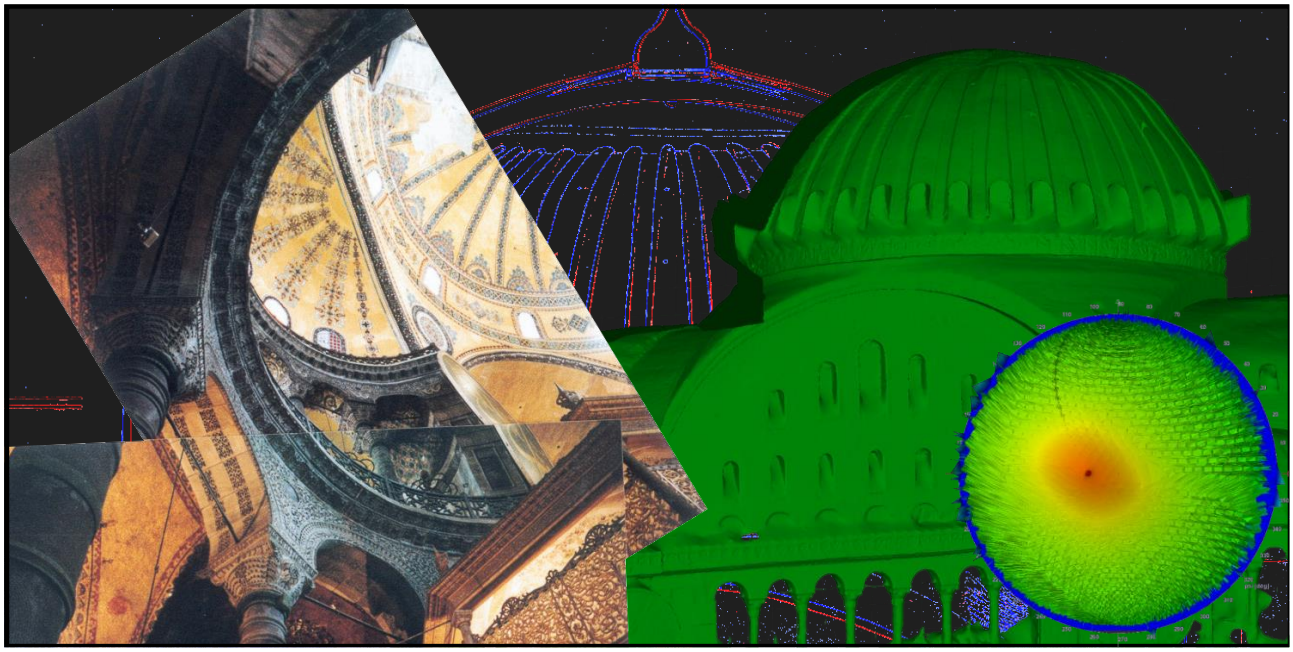

Fig. 2. Exemplary data from the Hagia Sophia project. Digital and analogue images from the photographic documentation (left). Digitised plans from a building survey (background). Triangular geometry derived from structure from motion techniques (right). Data from goniophotometric measurements (bottom right). © Authors.

The focus groups of the FID BAUdigital include research communities in the field of *built cultural heritage*. The FID is developed by the University Library Braunschweig, the University and State Library Darmstadt, the TIB – Leibniz Information Centre for Science and Technology and the Fraunhofer Information Centre for Planning and Building. It started in autumn 2020 and its current key activity is a community driven requirements analysis. The FID's goal is to provide web-services that *include AI applications*:

- Deep learning for image and video analysis, presumably based on the projects iART[7] and VIVA[8].
- Semantically enriched search with machine learning, presumably based on the project PubPharm (Wawrzinek et al., 2019).
- A Co-occurrent based recommendation system (Boubekki et al., 2017).

A case study for machine-readable data

In this context of research data management, the model of Hagia Sophia lends itself as a case to study the role of data, its citability and interoperability in cultural heritage. It allows to evaluate the internals of a dataset that presents a realistic, non-ideal case for curation. Approaches to transform

[3] https://www.dfg.de/en/research_funding/programmes/infrastructure/lis/funding_opportunities/specialised_info_services/index.html
[4] https://www.arthistoricum.net/en/about-us/
[5] https://propylaeum.de/en/about-us/
[6] https://www.fid-bau.de
[7] https://projects.tib.eu/en/iart/about/
[8] https://projects.tib.eu/en/viva/projekt/

the model from its current state are discussed as well as insights for the development of research data infrastructures and AI applications, including:

- extending research platforms in cultural heritage to comply to FAIR principles,
- utilising research data infrastructures and contribute to their development,
- how to link research to its sources and to comparable results,
- how to reach common data management solutions,
- how to benefit from AI applications.

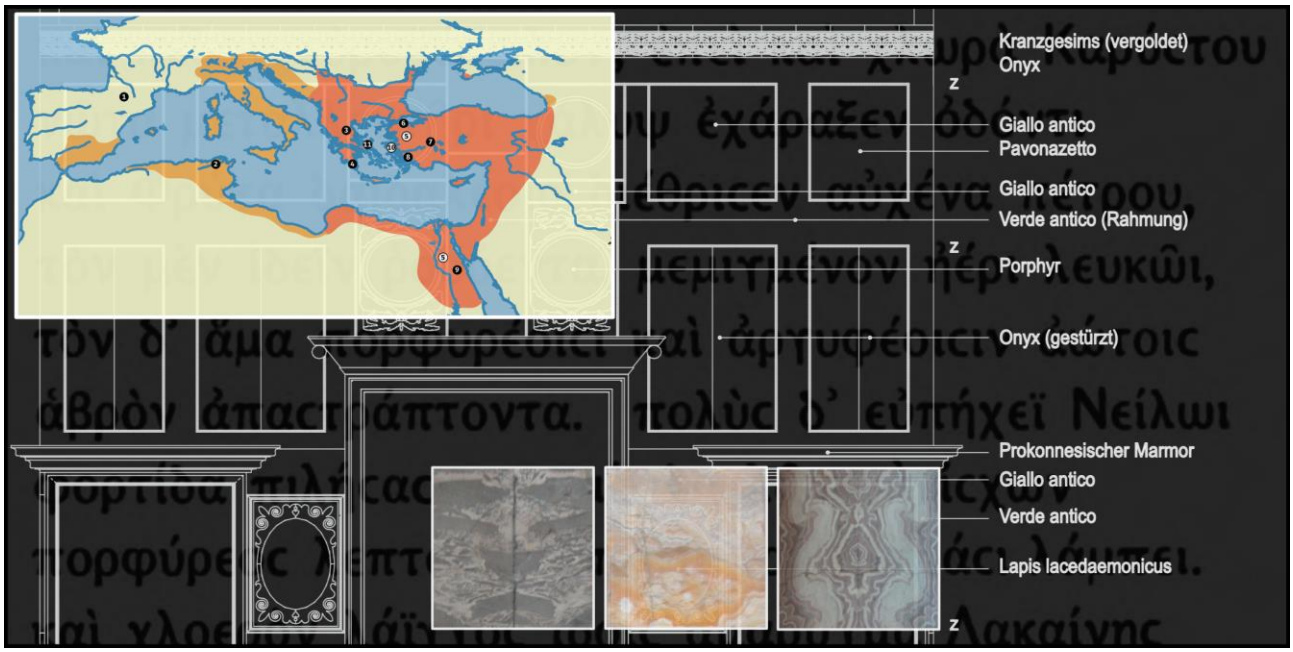

Fig. 3: External dependencies of the reconstruction model. The reconstructed marble decoration (background) describes names, colors, and origins of the materials. It is cross-referenced with present-day evidence in the building (bottom) and a marble collection related to Roman quarries. The model references these sources by page numbers of the catalogue but should rather allow open-linked data. © Authors.

The present heterogeneous dataset

The dataset is currently split into static content stored on a file server, and the version-controlled simulation environment. The former comprises a photographic documentation, digitised survey plans as well as photogrammetric data. It further includes measured optical properties, for example of Roman window glass, glass mosaic tesserae and marble samples (Fig. 2).

The reconstruction combines information from various sources to a plausible model (Fig. 3). For example, the wall decoration that fits the identified marble material from multiple imperial quarries forms a consistent geometry. The reconstruction refers to external sources, for example historic texts and objects from collections such as historic marble samples from Berlin. In this case it can only refer to pages of text (Veh, 1977) and figures in printed catalogues (Mielsch, 1985).

For the simulation environment simple, the choice of text-based triangular and polygonal mesh representations of all geometric objects has supported version control and has avoided compatibility problems over the decades. However, efficient editing of the reconstruction model required more complex entities due to the geometric constraints of the floor plan and the volumetric properties of the interior space. Therefore, editable CAD files of multiple proprietary formats were kept in parallel

to the simulation model. These binary files do not support versioning and separation into the model's tree-structure. Data has been lost due to conversion between different versions of the CAD software. The proprietary formats hindered a structured documentation of the reconstruction efforts.

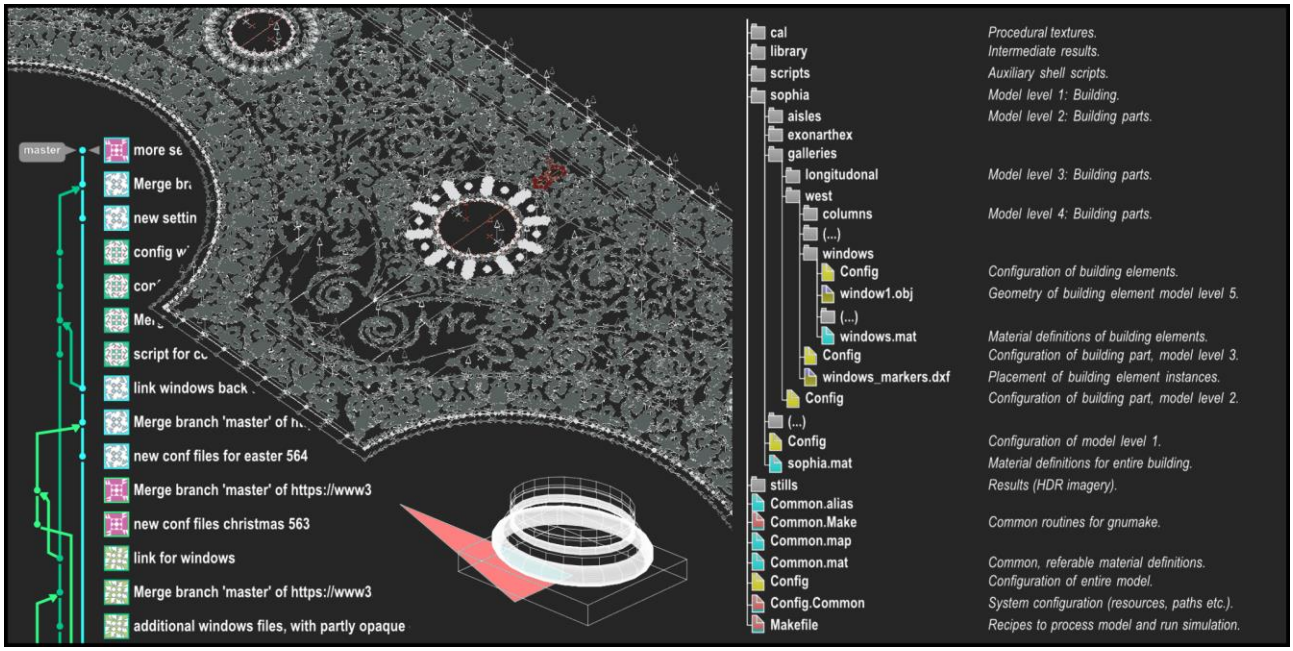

Fig. 4. The simulation environment combines the geometry of the reconstruction model (top left) with material models and configurations in a file system tree (right, Grobe et al., 2020). All data except images are stored in clear text formats allowing version control (left). Objects are placed as instances (bottom centre). © Authors.

The core of the simulation environment forms a directory tree that reflects the building's spatial structure (Fig. 4). Its geometry is separated into objects, stored as meshes (DXF and OBJ). To assemble the model, objects are instantiated, positioned, and oriented according to triangular placeholders stored in so-called marker files. Complex objects are represented by sub-trees. Reflection and transmission properties are stored as text-based parametrisations of the improved Ward-model (Geisler-Moroder and Dür, 2010), or as data-driven light scattering models in XML files (Ward et al., 2014). Non-uniform appearance and colour are represented by calibrated image files. Routines to process the model, and to start simulations are formulated as a dependency graph interpreted by GNU make. The simulations are further guided by viewpoints and time-steps. The entire simulation environment except the imagery is stored in text-files and under version control – initially by CVS, later SVN, and now git – that tracks the history of changes, including comments and contributors, combined with fragmented observations, sources, and records of important guiding considerations in text files within the directory tree.

The simulation results in numeric data. From that visualisations, mimicking the human perception of contrast; false-colour representations of photometric quantities such as the illuminance on surfaces; and diagrams are derived for publication (Fig. 5). These results are stored on the file server for internal use and disseminated through publications and on request by other researchers.

A concept with five data scopes for data management and AI application

In face of the heterogeneous research data and its development it is evident that no single approach fits all needs without developing a very specific application – and ontology – that lacks generality

and hinders access. Further examination of requirements for collaborative development, scientific transparency and documentation, machine-readable annotation, and dependencies to digital collections or external software-development lead to a separation into five data-scopes for the project data and digital publications as depicted in Figure 6: 1. Raw data collections, 2. Reconstruction and material models, 3. Simulation environment, 4. Simulation results and 5. Digital publications.

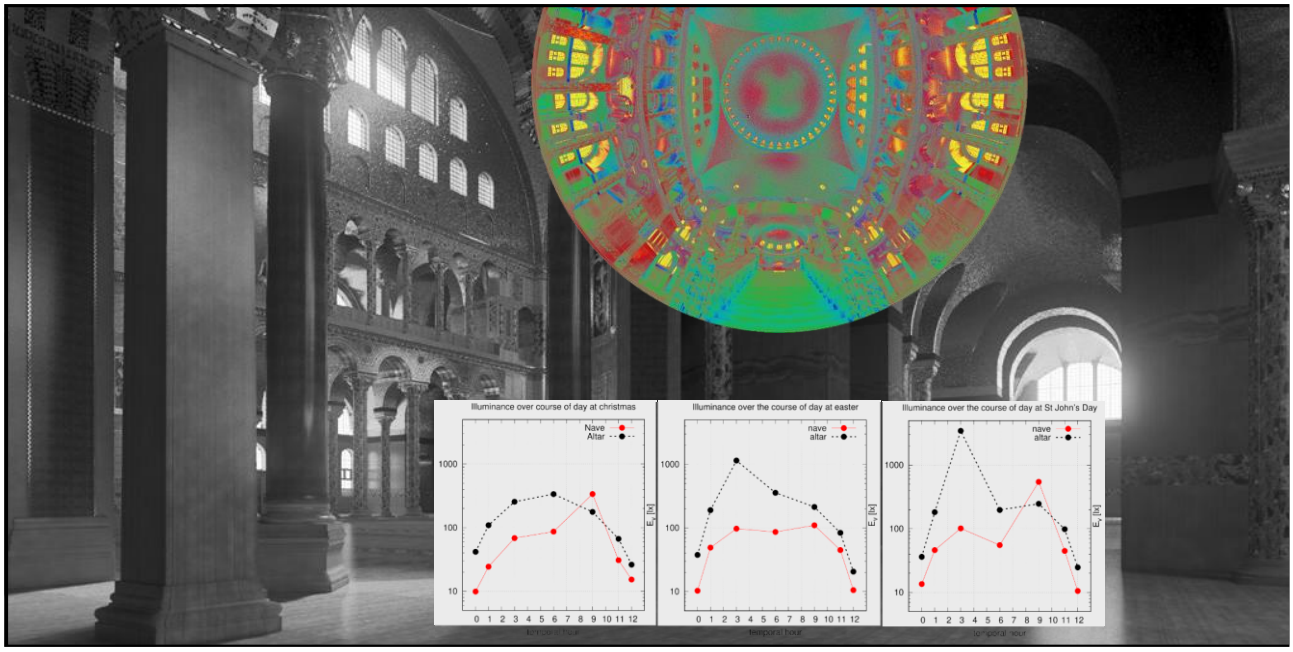

Fig. 5. Exemplary simulation results. Numeric data is stored in tabular form and visualised in diagrams (bottom) or false colour representation (top). Imagery visualises human perception models (background). © Authors.

This separation is compatible with the dataflow and dependencies within the project and is helpful for the integration into existing infrastructures or ontologies. It provides opportunities to find common solutions to similar problems in other projects and helps to define requirements for multi-purpose publication infrastructures and possible AI applications. For example, digital collections and object catalogues follow similar structures that can be reused for multiple purposes. They allow a standardised publication format as well as the application of AI assisted search tools.

References to objects in such structured data-models and suitable meta-data in all publications will enhance these possibilities and provide additional transparency. Internal and external linking between objects in the different data-scopes enables knowledge graphs as a form of scientific text, and defines anchor points for search infrastructures. For example, the photographic documentation could be ordered with AI assistance for spatial reference, search and relation to other image sources. Finally, the separation allows a step-by-step approach for the demanded publication of all research objects.

Conclusions

The proposed data scopes form horizontal layers in the diagram in which the output of a project on the right relates to the input on the left (Fig. 6). The separation presents opportunities for reuse and AI application. The latter may help with search and data mining in texts and digital collections, e.g. through word2vec (Mikolov et al., 2013), word movers' distance (Kusner et al., 2015) or a co-occurrent based recommendation system to find comparable results and publications (Boubekki et al.,

2017). Another potential opportunity would be the organisation of photographic documentation for spatial reference, search and relation to other image sources. Furthermore, custom AI applications may address particular research questions, for example to solve the fitting parameters of material models or to enhance simulation and data analysis. AI applications are part of and depend on a wider complex of management and publication of research data that link individual projects in a wider community. Segmenting project data into scopes is hoped to foster a wider use of AI applications.

With the goal to enhance the ability of machines to automatically find and use data, the presented concept provides opportunities for a more general discussion about data management and exchange in similar cultural heritage projects, the application of the FAIR Principles, and the further development of collaboration platforms such as the aforementioned Specialised Information Services.

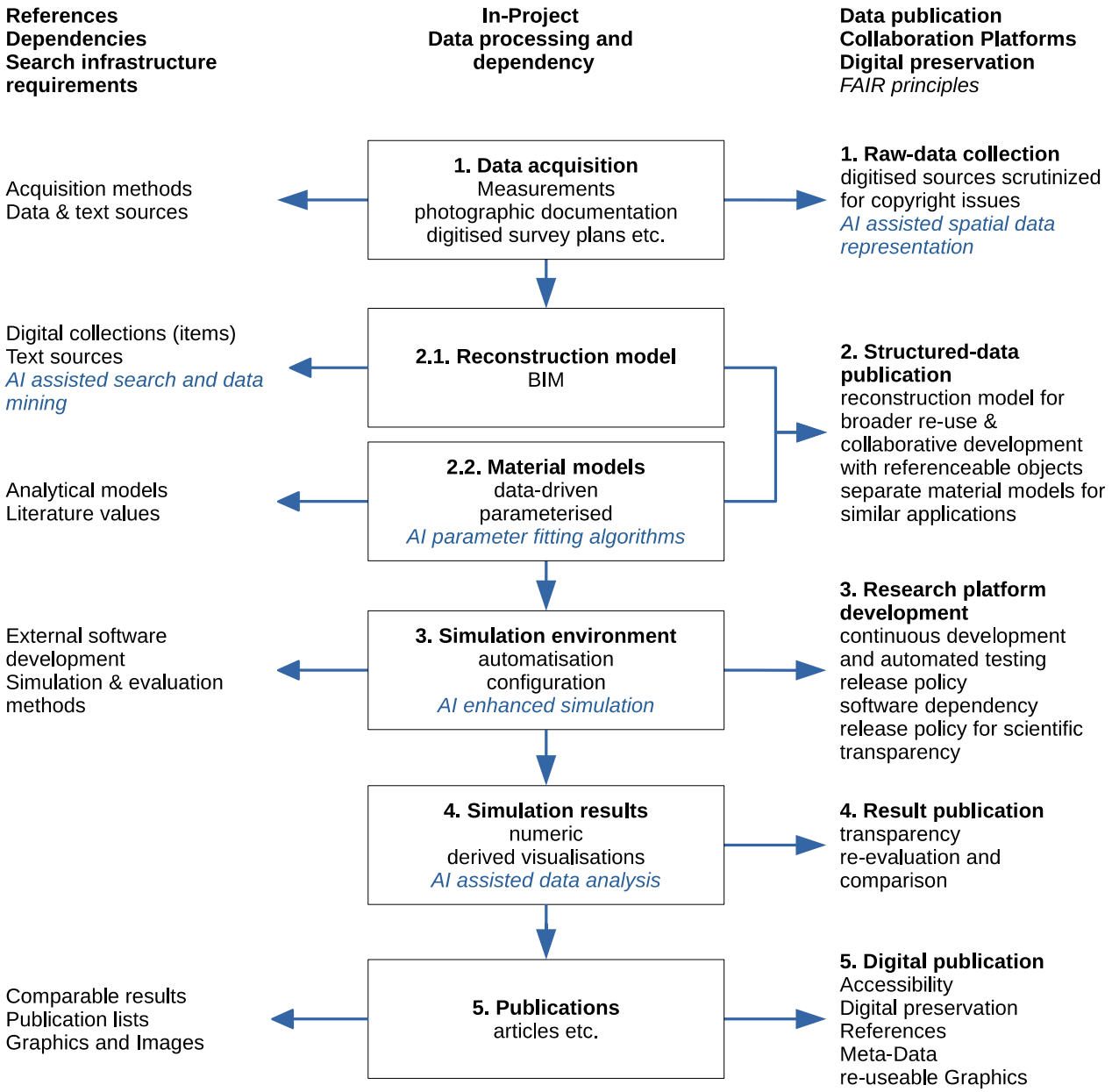

Fig. 6: Proposed five scopes for in-project data-processing, references, publication and AI applications. © Authors.

Acknowledgements

The discussed model of Hagia Sophia was developed by R. Stichel, H. Svenshon, O. Hauck, the authors, and others. It was further extended as part of the first author's Ph.D. at Technische Universität Darmstadt.

Funding

The discussed model of Hagia Sophia was developed with the support of Deutsche Forschungsgemeinschaft (DFG #5194526), the Fritz Thyssen Foundation (Az. 20.18.0.00 AA). This work on this paper is founded by the Deutsche Forschungsgemeinschaft as part of the project "FID Bauingenieurwesen, Architektur und Urbanistik digital (BAUdigital)" (#432774435).

Conflict of Interests Disclosure

The authors declare no conflict of interests.

Author Contributions

Both authors contributed equally to the paper.

References

Boubekki, A., Brefeld, U., Lucchesi, C. L. and Stille, W. (2017). 'Propagating Maximum Capacities for Recommendation' in Kern-Isberner, G., Fürnkranz, J. and Thimm, M. (eds.) KI 2017: *Advances in Artificial Intelligence*. Cham: Springer, pp. 72–84.

Geisler-Moroder, D. and Dür, A. (2010). 'A new Ward BRDF model with bounded albedo'. *Computer Graphics Forum*, 29, pp. 1391–1398.

Grobe, L. O., Noback, A. and Inanici, M. (2020). 'Challenges in the simulation of the daylight distribution in late antique Hagia Sophia', in Diker, H. F., Esmer, M. and Dural, M. (eds.) *Proceedings of the International Hagia Sophia Symposium*. Istanbul: Fatih Sultan Mehmet Vakif University Publications, pp. 661–685.

Hauck, O., Noback, A. and Grobe, L. (2013). 'Computing the "Holy Wisdom"', in Hans Bock, G., Jäger, W. and Winckler, M. J. (eds.) *Scientific Computing and Cultural Heritage*. Heidelberg: Springer, pp. 205–216.

Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. (2015). From word embeddings to document distances. In F. R. Bach and D. M. Blei (eds.) *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France, pp. 957–966.

Mielsch, H. (1985). *Buntmarmore aus Rom im Antikenmuseum Berlin*. Berlin: Staatliche Museen Preussischer Kulturbesitz.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). 'Distributed representations of words and phrases and their compositionality'. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2. Red Hook, NY, USA: Curran Associates Inc., pp. 3111–3119.

Noback, A., Grobe, L. O. and Inanici, M. (2020). 'Hagia Sophia's sixth century daylighing', in Diker, H. F., Esmer, M. and Dural, M. (eds.) *Proceedings of the International Hagia Sophia Symposium*. Istanbul: Fatih Sultan Mehmet Vakif University Publications, pp. 687–706.

Veh, O. (ed.) (1977) *Prokop: Die Bauten*. München: Heimeran.

Ward, G., Kurt, M., Bonneel, N. (2014). 'Reducing anisotropic BSDF measurement to common practice', in *Eurographics 2014 Workshop on Material Appearance Modeling: Issues and Acquisition*. Lyon, France, pp. 5–8.

Wawrzinek, J., González Pinto, J. M. and Balke, W. (2019). 'Linking Semantic Fingerprints of Literature – from Simple Neural Embeddings Towards Contextualized Pharmaceutical Networks', in Doucet, A., Isaac, A., Golub, K., Aalberg, T. and Jatowt, A. (eds.) *Digital Libraries for Open Knowledge*. Cham: Springer, pp. 33–40.

Wilkinson, Mark D. et al. (2016). 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific data*, 3. doi:10.1038/sdata.2016.18.