From me to you: peer-to-peer collaboration with linked data

Jan Kaßel and Dr. Thomas Köntges

Leipzig University, Germany

Corresponding author: Jan Kaßel, jan.kassel@studserv.uni-leipzig.de

Abstract

In recent years, Digital Humanities' collaborative nature has caused an awakening of digitally native research practice, where interdisciplinary workflows commonly feed into centralized data repositories. Connecting these repositories, the W3C's Web Annotation specification builds upon linked data principles for targeting any web resource or linked data entity with syntactic and semantic annotation. However, today's platform-centric infrastructure diminishes the distinction between institutions' and individuals' data. This poses issues of digital ownership, interoperability, and the privacy of data stored on centralized services. With Hyperwell, we aim to address these issues by introducing a novel architecture that offers real-time, distributed synchronization of web annotations, leveraging contemporary peer-to-peer technology. Extending the peer-to-peer network, institutions provide Hyperwell gateways that bridge peers' annotations and the web. These gateways affirm a researcher's affiliation, acting as a mere mirror of that researcher's data, while maintaining digital ownership.

Keywords

linked data; peer-to-peer systems; annotation; real-time collaboration

INTRODUCTION

Research projects in the Digital Humanities (DH) continually establish new standards in digital workflows with novel tools. This development is facilitated by an increasing number of open-source projects, available for free use and maintained by a large community of voluntary contributors. Commonly respected guidelines for managing data in research projects, such as the [FAIR] principles (Findable, Accessible, Interoperable, Reusable), establish a theoretical framework for these workflows, complementing technical architecture. Furthermore, by incorporating end-to-end, persistent references with technologies such as Digital Object Identifiers (DOIs), contemporary digital publishing takes on these frameworks. Thereby, references to particular versions of data can be resolved via open archives such as [Zenodo] and [HAL].

Reaching this level of quality in digital workflows is a major achievement for academia. Yet, the way we currently treat our data repositories is, in general, negligent. Often archived on platforms running on remote hardware in far-away data centers, people are rarely able to *exactly* prove (that is, mathematically) *when* and *what* they contributed to a data repository. After all, at this point, digital data ownership is mostly given when data resides on one's personal computing device. Local availability of data is itself advantageous: [Kleppmann et al., 2019] coined the term *local-first applications* in reference to software that operates on locally stored data, as opposed to the popular approach of *thin clients*, where data is mainly pushed to servers running remotely. However, local-first applications further question the necessity of centralized services for real-time collaboration—take Google Docs—by asserting that even as data leaves one's devices, digital ownership can be proven by cryptographic means. This concept provides one major takeaway: personal data doesn't necessarily need to be shared with third parties when collaborating with peers.

Applying this concept to research, we propose a distinction between personal and institutional data in the context of annotation. Artifacts belonging to the institutional domain, such as classic texts or digitized manuscripts, are stored remotely. Annotations on these sources, however, can be stored on an individual's device. Our work describes such workflows and, by introducing a system called Hyperwell, provides an architecture that leverages contemporary peer-to-peer technology to realize local-first, distributed, and collaborative annotation environments.

I COLLABORATION IN THE DIGITAL HUMANITIES

In the wake of Web 2.0—a technological development that emerged after the dot-com bubble burst in 2001—the internet became more open and accessible to the general public. Collaboration in research and particularly in DH adopted this trend, according to [Davidson, 2008], while the web tended towards the social sharing of resources on websites. Two advancements then materialized within DH: as human expression became increasingly digital, the humanities' workflows turned digital, too; subsequently, information on the past has been digitized, emphasizing digital practice within the humanities.

1.1 Exposing history

The FAIR principles consist of four imperatives for digital workflows—Findability, Accessibility, Interoperability, and Reusability—and have been rapidly adopted among scholars. Some best practices go even further by explicitly recommending version control and the recording of additional metadata, such as [Nowogrodzki, 2020]. Maintaining a log of changes is a convenient, self-documenting step in end-to-end digital workflows.

Over the years, the [Git] version-control system has been established as a popular tool in software development, as it allows users to write code in a change-based workflow. Collections of changes are bundled into a commit, which is then written into an immutable log and distributed to others. [Kreps, 2013] stated that this append-only log makes it straightforward for log-based databases to maintain a persistent history of changes and enables time-traveling between past states.

1.2 Web annotation

With pen-and-paper annotation, one's expression is basically unlimited: lines, highlights, and words can connect any visible artifact. Transforming such free interaction into digital processes can entail an unforeseeable complexity. However, as [Marshall, 1997] observed while analyzing a collection of students' used textbooks, a majority of their annotations fit into a categorization of six distinct schemes, varying in their focus and level of detail from simple marginal symbols up to complex sidenotes.

Published by the World Wide Web Consortium (W3C), the Web Annotation specification by [Sanderson et al., 2013] is built upon an ontology that is expected to be as versatile as possible, covering annotation schemes such as the above by providing varieties of target selection mechanisms and content types. The fundamental components of an annotation within this ontology are its target (a web resource) and its body (an entity). By using semantic expressions, an annotation body can convey complex structures, like their own motivation (for example, editing a passage) and their relations to other objects (for example, referring to a particular person). These structures could meet the semantic requirements of digital datasets in the humanities, such as digital gazetteers.

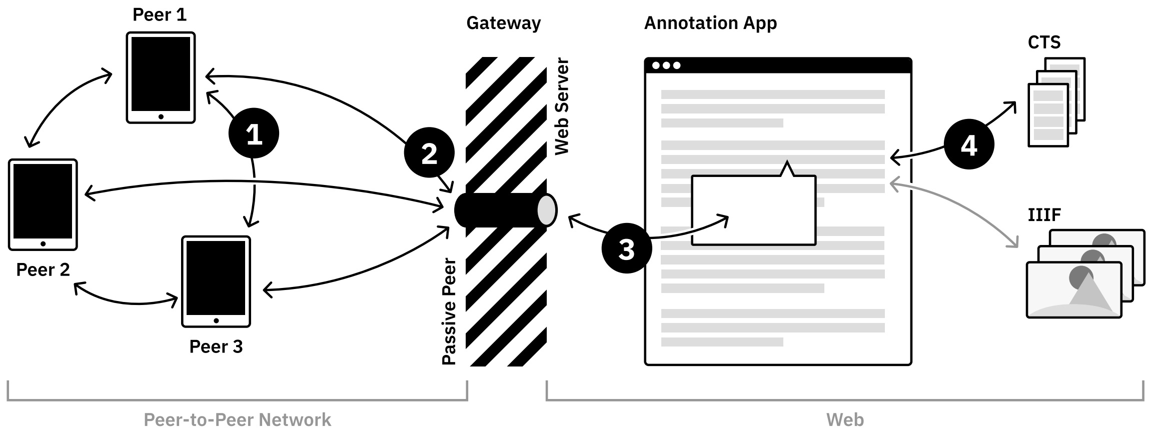

Figure 1: The Hyperwell architecture. Peers exchange their notebooks in real-time (1). Gateways, run by institutions, archive selected notebooks and bridge them into the web (2). Web applications can access annotations via gateways, as they implement the Web Annotation protocol (3). These applications can load canonical resources via services such as CTS or IIIF (4).

II PERSONAL ANNOTATIONS WITH HYPERWELL

With our work on personal annotation, Hyperwell, we aim to address several concerns at once. Hyperwell imposes a clear architectural distinction between *personal* and *institutional* work enabled by leveraging peer-to-peer technology. We introduce the notion of a digital, personal notebook, residing on individuals' devices. Such a notebook is a collection of annotations on a single resource and complies with the [Web Annotation Data Model].

Figure 1 outlines the architecture of Hyperwell. Each notebook can be shared directly with selected peers. By exposing an immutable, append-only log of all changes made to it, each version of a notebook can be referenced individually and facilitates archiving. Special requirements on the data types used allow the realization of real-time collaboration without the need for a central authority to resolve merge conflicts between multiple peers. Finally, end-to-end referencing of such workflows becomes possible, as each version of a notebook can target canonical resources, such as passages via Canonical Text Services (CTS) or multimedia resources via [IIIF].

2.1 Bridging the gap with gateways

Traditional notions of *pure* P2P networks impose the requirement that within such networks, resources are shared by nodes homogeneously, as [Schollmeier, 2002] has noted. This homogeneity ensures that data within these networks is genuinely decentralized. System architects occasionally break with this homogeneity for various reasons; for instance, to improve performance or to ensure interoperability with other networks.

In a patent, [Matsubara and Miki, 2010] sketched how such interoperability could be realized by describing a gateway between a peer-to-peer network and the web. Similar to the gateway pictured in Figure 1, it acts as an intermediary between individual peers and requests from web clients. [Kaßel, 2020] implemented a gateway for the Hyperwell architecture that acts both as a passive peer within the peer-to-peer network and as a server for web clients. The gateway implements the [Web Annotation Protocol] for compliance with environments using the Web Annotation specification.

2.2 Institutional archival of notebooks

Along with the aforementioned advantages of peer-to-peer networks, there is also an associated risk: if the device serving a certain resource is disconnected from the network, this resource is not available to other peers. In an academic context, however, and especially for published datasets or collaborative work, continuous data availability is crucial.

Supporting nodes can improve data availability and, hence, the overall quality of a peer-to-peer network. By *pinning* or *seeding* resources, such nodes can mirror other peers' data without actually causing any increase in the centralization of resources; on the contrary, this affects a further distribution. Furthermore, as volatile infrastructure, supporting nodes maintain original ownership. Seeding became an integral part of BitTorrent systems, as [Legout et al., 2007] have noted, while incentive mechanisms can further help to maintain supporting infrastructures. Within the architecture of Hyperwell, such a supporting infrastructure operates in the interest of institutions: they can ensure data availability for affiliated researchers as well as archive repositories in their entirety, as each repository contains its complete history.

2.3 Real-time collaboration

Merging changes from different sources is rarely a trivial task, especially within a distributed environment. Collaborative applications like Google Docs and Trello use a central authority—their servers—to solve these conflicts, but this requires a continuous internet connection from clients and exclusively relies on this authority. However, contemporary technology can already address this challenge: Conflict-Free Replicated Data Types (CRDTs) impose several theoretical merging strategies on data in order to obtain consistent and fail-safe merging in distributed networks, which [Kleppmann et al., 2019] have leveraged for decentralized collaboration "in spite of the cloud".

To realize this functionality in Hyperwell, we have utilized [hypermerge], a library that combines several technologies of distributed computing for realizing real-time, peer-to-peer collaboration:

- [automerge], a JavaScript-based CRDT implementation,
- [hypercore], a distributed append-only log, allowing us to immutably store and distribute changes on a dataset, and
- [hyperswarm], a stack of distributed networking technologies for establishing connections between peers without centralized discovery.

Both hypercore and hyperswarm are being actively developed by the Dat Protocol foundation. This not-for-profit organization builds the Dat data-sharing protocol, introduced by [Robinson et al., 2018], for use in research and civic technology.

III OUTLOOK

With Hyperwell, we have produced a collaborative, peer-to-peer annotation system that builds upon contemporary peer-to-peer technology and implements the Web Annotation specification. Hyperwell aims to address issues concerning data privacy and digital sovereignty in DH research, facilitated by centralized platforms. Establishing a separation of personal and institutional data, we present a gateway service for Hyperwell, bridging personal, distributed annotation notebooks and the web for interoperability with existing annotation environments.

A forthcoming Master's thesis by Jan Kaßel at Universität Leipzig includes further work that will focus on evaluating the presented architecture and its entailing collaborative workflows. To provide a necessary prototype environment, we aim to build two applications: First, a local-first notebook application that will manage annotations by storing them locally as well as indexing them for local search. Second, for conducting actual usability and workflow evaluation, a proof-of-concept collaborative environment that integrates canonical resources, such as CTS passages or multimedia IIIF manifests.

References

automerge. https://github.com/automerge/automerge.

Davidson C. N. Humanities 2.0: Promise, perils, predictions. *PMLA*, 123(3):707–717, 2008. doi: 10.1632/pmla.2008.123.3.707.

FAIR. https://www.go-fair.org/fair-principles/.

Git. https://git-scm.com/.

HAL. https://hal.archives-ouvertes.fr/.

hypercore. https://github.com/mafintosh/hypercore.

hypermerge. https://github.com/automerge/hypermerge.

hyperswarm. https://github.com/hyperswarm/hyperswarm.

IIIF. https://iiif.io/.

Kaßel J. Hyperwell gateway, Zenodo, 2020. doi: 10.5281/zenodo.3631524.

Kleppmann M., Wiggins A., van Hardenberg P., and McGranaghan M. Local-first software: you own your data, in spite of the cloud. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software - Onward! 2019*, pages 154–178. ACM Press, 2019. doi: 10.1145/3359591.3359737.

Kreps J. The log: What every software engineer should know about real-time data's unifying abstraction, 2013. https://engineering.linkedin.com/distributed-systems/log-what-every-software-engineer-should-know-about-real-time-datas-unifying.

Legout A., Liogkas N., Kohler E., and Zhang L. Clustering and sharing incentives in BitTorrent systems. *ACM SIGMETRICS Performance Evaluation Review*, 35(1):301, 2007. doi: 10.1145/1269899.1254919.

Marshall C. C. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries - DL '97*, pages 131–140. ACM Press, 1997. doi: 10.1145/263690.263806.

Matsubara D. and Miki K. Method and apparatus for peer-to peer access, 2010. https://patents.google.com/patent/US7769881B2/en.

Nowogrodzki A. Eleven tips for working with large data sets. *Nature*, 577(7790):439–440, 2020. doi: 10.1038/d41586-020-00062-z.

Robinson D. C., Hand J. A., Madsen M. B., and McKelvey K. R. The dat project, an open and decentralized research data tool. *Scientific Data*, 5:180221, 2018. doi: 10.1038/sdata.2018.221.

Sanderson R., Ciccarese P., and Van de Sompel H. Designing the W3C open annotation data model. In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 366–375. ACM Press, 2013. doi: 10.1145/2464464.2464474.

Schollmeier R. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In *Proceedings First International Conference on Peer-to-Peer Computing*, pages 101–102. IEEE Comput. Soc, 2002. doi: 10.1109/P2P.2001.990434.

Web Annotation Data Model. https://www.w3.org/TR/annotation-model/.

Web Annotation Protocol. https://www.w3.org/TR/annotation-protocol/.

Zenodo. https://zenodo.org/.