Scholarly Social Machines: A Web Science Perspective on our Knowledge Infrastructure

David De Roure
david.deroure@oerc.ox.ac.uk
Oxford e-Research Centre, University of Oxford, UK
The Alan Turing Institute, London, UK

Pip Willcox
pip.willcox@nationalarchives.gov.uk
The National Archives, London, UK

ABSTRACT

A Knowledge Infrastructure comprises the people, artefacts, and institutions that generate, share, and maintain knowledge, very often mediated by the Web. Our scholarly Knowledge Infrastructure is evolving as researchers embrace digital techniques enabled by increasing availability of digital data, computational power, and analytical tools and techniques. Crucially, the social structures are changing also. Taking a Web Science approach, this paper encourages the reader to view the scholarly Knowledge Infrastructure as an ecosystem of interacting and evolving *Social Machines*. We illustrate these *Scholarly Social Machines* with a series of descriptive examples, and reflect on these to propose *Scholarly Primitives* associated with Scholarly Social Machines. We suggest that this approach facilitates a holistic understanding of our scholarly Knowledge Infrastructure and informs its evolution.

CCS CONCEPTS

• **Information systems → Crowdsourcing**; *Social networks*; **Digital libraries and archives**; • **Applied computing** → Arts and humanities.

KEYWORDS

Social Machines, Knowledge Infrastructure, Digital Humanities, e-Research, Crowdsourcing, Scholarly Primitives

ACM Reference Format:
David De Roure and Pip Willcox. 2020. Scholarly Social Machines: A Web Science Perspective on our Knowledge Infrastructure. In *12th ACM Conference on Web Science (WebSci '20), July 06–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3394231.3397915

1 INTRODUCTION

Knowledge Infrastructures have been defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds", and include individuals, organisations, routines, shared norms and practices. [13]. We find it useful to adopt this inclusive and comprehensive definition, which for example would include the Web Science conference, an archive or a library, and importantly includes people: it is fundamentally sociotechnical.

Edwards also observes "Infrastructures are not systems, in the sense of fully coherent, deliberately engineered, end-to-end processes. Rather, infrastructures are ecologies or complex adaptive systems; they consist of numerous systems, each with unique origins and goals, which are made to interoperate by means of standards, socket layers, social practices, norms, and individual behaviors that smooth out the connections among them."

Our scholarly Knowledge Infrastructure is evolving as researchers embrace digital techniques to explore established research fields through new ways and envisage new terrain, afforded by digitised and born-digital collections and the wealth of data emitted from our increasingly digital lives. Pervasive adoption of technology, coupled with the co-creation of new social processes, has created a new and complex space for scholarship—in which, for example, citizens both generate and analyse data as they interact at the intersection of the physical and digital. This rapid change in elements of our infrastructure is sometimes in tension with its established forms already, and from here we can anticipate more computational power, more automation, and further adoption of Artificial Intelligence.

How then do we make sense of such systems, describe and understand them, and plan for a future that enables useful, subversive and creative responses to what Edwards describes as "disarray and disjunction"? The authors have been responsible for digital scholarship across a spectrum of disciplines and are keenly aware of innovation in the methods and the infrastructure that supports such scholarship. We are also aware of the challenge of stepping back to consider the scholarly Knowledge Infrastructure holistically: every researcher is part of this system and can see it most easily from an individual perspective.

As in Web Science, we must step back and study the ecology as a whole—the Web, or the Knowledge Infrastructure, as an evolving artefact: the cover artwork on [13], viewing Knowledge Infrastructure from the outside as a complex artefact, captures this notion well. Our approach is to view Knowledge Infrastructures through the lens of *Social Machines*, a perspective that has attracted particular attention in the Web Science community. It is also informed by a body of work in the social sciences, for example Meyer and Schroeder's *Knowledge Machines* [15]. We then ask what this tells us about our *Scholarly Primitives*—the methods common to scholarly activity across disciplines and hence mutually reflected in the Knowledge Infrastructure.

We introduce the concept of Social Machines in Section 2. We then apply this lens in Section 3 to a series of examples of scholarly Knowledge Infrastructure which have attracted study. Section 4

WebSci '20, July 06–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397915

develops this further by taking a longer view, looking at established Knowledge Infrastructure and decoding it as Social Machines which predate the Web, then in Section 5 we glance into a creative, automated future. Finally in Section 6 we reflect on all these machines from the perspective of Scholarly Primitives, from a holistic vantage point. All the examples draw on the authors' direct experience.

2 SOCIAL MACHINES AND THE ECOSYSTEM PERSPECTIVE

Emerging scholarly methods embrace citizen engagement with the digital world at scale, together with increasing computational capability. This scaling-up has necessitated increasing automation, and more recently the adoption of machine learning techniques to build on human effort in order to process content at scale. This is the complex space which is described variously as citizen research or science, crowdsourcing, social computing, open innovation, collective intelligence and human computation; it can be characterised as having both high computational complexity and high collaboration complexity [15]. Web Science is a means of studying it, and one of our motivations has been to explore whether there might be an underlying abstraction, model, pattern or framework which assists in making sense of this space.

For this we turn to the concept of Social Machines, which has become an established lens to describe the sociotechnical systems of Web Science. Berners-Lee and Fischetti's definition of Social Machines talks of the stage set for "an evolutionary growth of new social engines" [2]. This is not about people in service of machines, but about the digitally empowered citizen: "The ability to create new forms of social process would be given to the world at large, and development would be rapid". This prediction anticipated social media such as Twitter, a place where citizens are empowered to create new social process. We suggest that the lens of the Social Machine is useful in describing the trajectory of sociotechnical systems [18] and hence to scholarship.

The authors have reported previously on the study of Social Machines in which we have taken the perspective of the *ecosystem* of Social Machines, rather than focusing on individual machines, so that we can also understand their coupling together [9, 10]. For example, we see coupling via shared artefacts, and via people: by making it explicit that people are using multiple machines we can discuss how they choose to assign their attention between them, and describe their journeys [16]. We do this by considering the processes they are enacting, an approach we rehearsed in our prior work on the application of the Social Machines lens to a location-based online augmented reality game [8], where we imagined the pseudocode that a human could be 'executing'.

3 SCHOLARLY KNOWLEDGE INFRASTRUCTURES

We define *Scholarly Social Machines* to be the Social Machines in our Knowledge Infrastructures. In this section we present an illustrative range of Scholarly Social Machines, and for each we identify the significant characteristics of the machines which inform our subsequent reflections on Scholarly Primitives.

3.1 Wikipedia

Wikipedia is a powerful example of a Knowledge Infrastructure and Scholarly Social Machine. It is the most widely used general reference work on the Web, growing since 2001 to hold over 6 million articles in the English language (for comparison, *Encyclopaedia Britannica* has less than a million). To situate this against the pre-digital infrastructure, a print edition of Wikipedia in 2015 comprised over 7000 volumes of 700 pages each [14].

The volunteer content creation and editorial processes of Wikipedia make it distinctive with respect to other encyclopedias and scholarly electronic editions. It is an example of what Siemens calls a *social edition*, capitalising on engaged knowledge communities inside and outside the academy [21].

As a Social Machine, Wikipedia is clearly operating at scale and 'on the web', and the processes that run it have emerged through community interactions over nearly two decades, such that its behaviour can be described as socially constituted. This is described in its self-definition: "Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteer editors using a wiki-based editing system." It is an open platform and has been widely studied including critical analysis of its growth, questioning for example whether it resists new content and deters newcomers [24].

As an example of a Scholarly Social Machine, Wikipedia is a sustained open platform in which editorial and administrative processes have evolved to maintain the quality of the content. It is interesting to note that the scale is today handled by bots, which automate various behind-the-scenes curation tasks including cataloging and dealing with vandalism. Bots have also been used to initiate content creation by generating stub articles automatically based on external sources. The rich set of approved bots curating Wikipedia demonstrates another form of contribution [28], and a model for automation in the Knowledge Infrastructure.

3.2 Zooniverse

The Zooniverse citizen research, or 'people powered research', platform has evolved through several versions and today reports around 500 million classifications by 2 million registered volunteers. Starting out in 2007 as Galaxy Zoo, the platform now delivers citizen research projects across multiple domains—a significant Knowledge Infrastructure capability in itself which, like Wikipedia, has attracted study [25]. Its current evolution features the Zooniverse Project Builder, enabling anyone to create a citizen research project. The tools it provides include annotation (tagging, drawing) and transcription. What started out as the Social Machines of Galaxy Zoo has evolved into a platform for accessible Social Machines creation.

Today the original Galaxy Zoo project still exists on the Zooniverse platform, and with an option to use an enhanced version. At the time Galazy Zoo was originally conceived, there were too many images for one person to classify and the online platform was developed to enable a larger volunteer community to help. The size of the volunteer workforce cannot keep pace with the increasing supply of images, so the platform has now turned to machine learning. Here is the explanation from the Zooniverse site [29]:

> In an effort to speed up classifications to cope with the large number of galaxies we expect to receive from new surveys, we've been working on ways to combine your classifications with those of machines, inspired by the idea that the combination of both automatic and human classification may be more powerful than either alone. If you choose the 'Enhanced' work flow, you will be much more likely to see the top 100 galaxies our galaxy-classifying robot thinks it needs help with in order to improve. All galaxies will be seen by at least a few volunteers to make sure we aren't missing anything. If you'd rather just see a random selection of available galaxies, choose 'Classic'.

Zooniverse is a sustained and evolving platform for Social Machines [22] (perhaps even a Social Machine for creating Social Machines). The inclusion of discussion fora, enabling the citizens to discuss their annotations and practice with each other and projects' core team of researchers, has been significant in terms of engagement and scientific outcomes, contrasting with crowdsourcing solutions based on microtasks that are conducted independently. Today it also exemplifies the adoption of machine learning alongside volunteers' effort to contend with the increasing scale of content.

3.3 myExperiment

myExperiment [7] is a workflow commons, launched in 2007. It was itself an experiment in creating a social website for sharing scientific (computational) workflows, deliberately constructed according to the Web 2.0 design principles [17]. A niche piece of Knowledge Infrastructure, the site is active today with some 4,000 workflows.

myExperiment gave a glimpse of the future of scholarly communication. Conceived for an era of Open Science, Open Source Software and Reproducible Research, the project also defined the notion of the *Research Object*, a new kind of artefact for sharing within the scholarly communication ecosystem [1]. Additionally the focus on computational workflows gives some glimpses of a more computationally enabled future, with *in silico* experimentation, automated curation, and executable documents [6]. 10 years on these insights are being validated, as we see code running remotely over data ("non-consumptive research"), Jupyter notebooks, computational archival science, and increasing adoption of AI.

An episode in the history of myExperiment brings an example of multiple interacting Social Machines within the ecosystem, and hence can present a focus for the Social Machines lens. The website was protected by reCAPTCHA to ensure only humans could create accounts—this was in itself a Social Machine, because reCAPTCHA at that time was using humans to transcribe digitised text. A spate of spam accounts on myExperiment turned out to be the result of a website which rewarded people for creating accounts to place products and influence search rankings, itself a Social Machine. The response by the myExperiment administrators was to create a short-duration Social Machine of volunteers to identify and block the fake accounts, while the site was modified to make use of a blacklisting site during account creation—another Social Machine. reCAPTCHA today still identifies humans but asks people to identify objects in images rather than transcribe digitised text: the slogan has changed from "Stop Spam. Read Books." to "Easy on Humans, Hard on Bots".

myExperiment is an example of a Scholarly Social Machine that supports researchers who are using automation, sharing process rather than data. It is also an example of a specialist Knowledge Infrastructure which was designed to couple readily with the larger ecosystem (so, for example, it supports some social network features but discussions occur elsewhere according to community practices). It is interesting to reflect on the relationship between "Web 2.0" and Social Machines, with clear synergies: the "long tail", data-driven applications, users adding value, network effects, mashups, the perpetual beta, and a maxim of "cooperate, don't control" [17].

3.4 SALAMI

The *Structural Analysis of Large Amounts of Music Information* (SALAMI) project set out to analyse a large number of musical recordings [11, 23]. This one endeavour involved multiple Social Machines.

In the first phase, the analysis was crowdsourced to music graduate students to produce an annotated dataset which represented the *ground truth*. In contrast to many crowdsourcing examples which engage general volunteer communities, the participants were musically trained experts and were paid. This is the first Scholarly Social Machine in this example, and it produced a dataset which has been used in a number of subsequent projects. In the second phase, structural analysis software was used to analyse the music, and the results were compared with the ground truth in order to improve the outcomes. An interactive tool enabled exploration of the annotations from analysis both human and machine.

This structural analysis software was itself the product of a Social Machine. The Music Information Retrieval (MIR) community comes together in an annual event called the *Music Information Retrieval Evaluation Exchange* (MIREX) [12], in which feature extraction algorithms (e.g. recognition of key, tempo, chords, genre, and structure) are evaluated against a reference corpus of digital music recordings. The features for each round are determined democratically within the community, and the results are published in the form of league tables. MIREX is our second Scholarly Social Machine in this example.

The structure of the project is shown in Figure 1. The outputs (SALAMI datasets, Linked Data) are reused by the scholarly community, hence connect into other Scholarly Social Machines. We note that Linked Data is itself an interesting example of a shared artefact which couples Social Machines and accumulates value through shared usage.

In contrast to the previous examples, SALAMI is a research project which created a short-duration and small-scale Social Machine (the grad-sourced analysis) which was "pop-up", paid and planned, while engaging productively with an existing sustained Social Machine (MIREX). It is an example of research practice today involving Social Machine design and creation.

4 SOCIAL MACHINES FOR PRINT AND PUBLISHING

Publishing is a crucial element of today's scholarly Knowledge Infrastructure, but it is not new. The publishing process now known as Open Science came about over 350 years ago with the publication of journals such as *Philosophical Transactions of the Royal Society*,

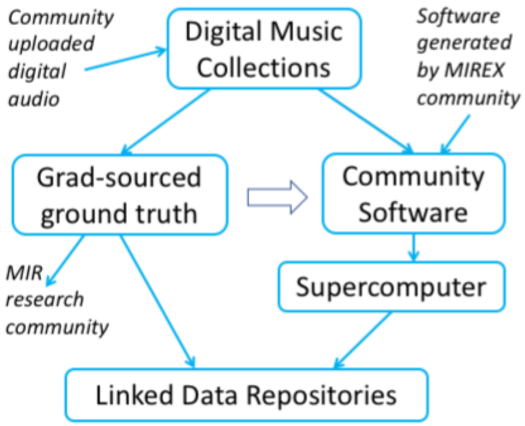
Community uploaded digital audio
Digital Music Collections
Software generated by MIREX community
Grad-sourced ground truth
Community Software
MIR research community
Supercomputer
Linked Data Repositories

Figure 1: Structure of the SALAMI project

a prestigious peer-reviewed open science journal still in publication today. Scholarly Communication has always been part of our Knowledge Infrastructure and it underpins the Social Machines of science itself. Recall for example the notion of 'standing on the shoulders of giants', attributed to Isaac Newton amongst others, including Robert Burton, who wrote: "a dwarfe standing on the shoulders of a Giant, may see farther then a Giant himselfe" [5].

As an exercise in applying the Scholarly Social Machines lens to an evolving infrastructure over centuries, in this section we relax the constraint that Social Machines are "on the Web" and take a long view on publishing. For this we choose one particular book, the Bodleian First Folio of Shakespeare's Plays [19]. There are many stages of co-creation in the story of this book, and here it acts as a probe into Knowledge Infrastructure over centuries. In line with our observational methodology, we present a summary of our work and then identify the Social Machines.

Current scholarship suggests that early modern plays were frequently co-authored. Authorship debates aside, plays written, as Shakespeare's were, for a company of actors by one of its members must be at the least influenced by their intended casting, if not co-created during rehearsal and performance. Furthermore, Shakespeare's plays, in the forms in which they reach us, are generally longer than could practically have been performed in contemporary theatres. It may not be unreasonable to suggest a performance was cut according to an anticipated audience's preferences, and not infeasible that versions of a play were co-created dynamically, by actors reacting to audience response.

The first collection of Shakespeare's plays (1623) is likely to derive in part from the prompt books—quintessentially socially constructed texts—of the King's Men (the acting company to which Shakespeare belonged). The First Folio, as it came to be known, was published as a joint venture by a consortium of printers—Edward Blount, William and later Isaac Jaggard, William Aspley, and John Smethwick—with two of Shakespeare's fellow actors and friends, John Heminge, and Henry Condell.

One copy of the First Folio was sent to the Bodleian Library in Oxford, presumably under its 1610 agreement with the Stationers' Company. This agreement, as well as supplementing a depleted library collection for the use of scholars at the cost to the library only of binding the books, laid the foundation for today's libraries of legal deposit—libraries as society's memory, not simply knowledge warehouses, but vital parts in the knowledge turn of creation, curation, reception, and inspiration of the Social Machine of scholarship, a Knowledge Infrastructure.

This copy left the library, probably sold after it had been superseded by the Third Folio of Shakespeare's plays of 1663/4 in an age before first editions were prized, and was lost to view for about 240 years. In 1905 Gladwyn Turbutt, an Oxford undergraduate, brought his family copy to the Bodleian Library's enquiry desk for advice on its dilapidated and lacklustre binding. It was identified as the Bodleian's long-lost copy by two librarians, Falconer Madan and Strickland Gibson, who, working with Turbutt, presented their findings at a Bibliographical Society meeting in a talk that was reviewed in *The Athenaeum* and *The Times*. Shortly afterwards an anonymous American offered to buy the book from the Turbutt family for £3,000—many times its market value. The wealthy American was later revealed to be Henry Clay Folger of Standard Oil, who was secretly collecting 'Shakespeareana' (Henry Clay and Emily Jordan Folger's Shakespeare Library was not founded until 1930). The desire to return the book to its original home inspired a private and then a public funding campaign—"Oxford men" (at whom the campaign was directed, although by neither education nor gender were the donors so restricted) contributing to the local and national commons.

The ultimately successful campaign saw the book returned to the Bodleian Library. Its fragile physical condition, from having been so much read in its early years on shelf in the library, meant access to the book was restricted and few scholars were able to study it. Emma Smith, Professor of Shakespeare Studies, mentioned this in a 2011 lecture, and inspired one of the authors to instigate a second public campaign, inviting blog posts on Shakespeare, and raising funds to stabilise, photograph, and publish a digital avatar of the book freely online in 2013 [4]. Further generous donations led to a digital edition of the book's text being published by a consortium from the University of Oxford (led by the Bodleian Libraries, with the e-Research Centre and IT Services) through a process of transcription and proofing, and encoding (compliant with the Text Encoding Initiative's Guidelines standard, Proposition 5, TEI p5), with images available through International Image Interoperability Framework (IIIF) technology [20]. The software is open source, and both text and images are published under a Creative Commons Attribution license, enabling their reuse in other Social Machines including research and education.

Having relaxed the 'on the Web' condition of the original definition, we identify multiple social-machine-like sociotechnical structures within the story of the First Folio, listed in Table 1. These are all examples of co-creation, where the processes themselves (for example, what today we call publication) were in early evolution.

5 SOCIAL MACHINES FOR CREATIVITY

Our final example brings together historical sources, AI and creativity, as we project our "long view" into the near future and extend to practice-based research.

Table 1: Social Machines in the history of the Bodleian First Folio of Shakespeare's Plays

1	Co-authoring	Early modern plays were frequently co-authored
2	Performance	Co-created during rehearsal or performance
3	Prompt books	Socially constructed texts
4	Publication	Printed as a joint venture by a consortium of printers
5	Copyright	Stationers' Company
6	Knowledge reception	Libraries
7	De-acquisition	Probably sold after it had been superseded
8	Enquiry desk	Identified as the Bodleian's long-lost copy
9	Scholarly community	Bibliographical Society meeting
10	The Press	Reviewed in *The Athenaeum* and *The Times*
11	Funding campaign (1905–6)	Private then public funding campaign
12	Funding campaign (2012)	Second public campaign
13	Publication (2013)	Digital avatar published freely online
14	Transcription, Proofing	Digital edition of the book's text produced
15	Encoding in standard	Text XML-encoded, compliant with Text Encoding Initiative Guidelines, Proposition 5
16	Publication (2014)	Digital edition of the book's text published
17	Re-use	Open source software, text, encoding schema and images Creative Commons Attribution license

Alter is a chamber ensemble work in which the human composer, Robert Laidlow, worked co-creatively with multiple AI systems [8]. *Alter* was performed in November 2019 as part of "Imagining the Analytical Engine", a musical tribute to Ada Lovelace, at the Barbican Centre in London, UK.

The text of *Alter* is written by an AI that audibly develops in coherence and philosophical scope. It learns from Ada Lovelace's correspondence, using a language model based on a 19th-century letter corpus supplied by the Electronic Enlightenment team at the Bodleian Libraries. It goes on to use an AI trained extensively on currently available modern English. Thus the narrative of the scene reflects the data science behind its production. The workflow is illustrated in Figure 2.

In this example, we draw on the Social Machines of 19th-century correspondence—including both the postal network and the scholarly network that used it. This in itself is an important area of study in Humanities, where digital methods are used to reassemble and interpret correspondence and knowledge networks. The two AIs' training data (from the 19th and 21st centuries) include correspondence so have the ability to mimic the form of these Social Machines. The music was co-created by the composer assisted by AI trained on previous work. We then have a Social Machine that relates to the previous example: performance with an audience, using both text and music that has been co-created with AI. Behind the scenes another Social Machine is in operation today: collective rights management makes sure that royalties are paid to musicians and composers when their work is performed. Ultimately the work produces a music score along with videos and recordings. These are the artefacts that then circulate in the Social Machines of contemporary music composition, performance, and research.

6 THE SCHOLARLY PRIMITIVES OF SCHOLARLY SOCIAL MACHINES

In the previous sections we have seen multiple Knowledge Infrastructure scenarios which include examples of Scholarly Social Machines. Here we reflect over these examples, with a particular aim to seek commonalities and patterns that might help identify a useful abstraction to assist with description and understanding, and to inform planning and interventions.

A particular inspiration for this analysis is the notion of *Scholarly Primitives.* These primitives:

> …refer to some basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation. These 'self-understood' functions form the basis for higher-level scholarly projects, arguments, statements, interpretations—in terms of our original, mathematical/philosophical analogy, axioms. [26]

The primitives are also intended to inform the infrastructure, which is particularly relevant to our Knowledge Infrastructure stance in this paper—noting of course that the research both influences, and is influenced by, the infrastructure.

The seven Scholarly Primitives suggested by Unsworth are:

Discovering **Annotating** **Comparing** **Referring**
Sampling **Illustrating** **Representing**

He also discusses **selection** (of, and in, content) and **linking** (annotation or association). We might expect these Scholarly Primitives to be applicable to Scholarly Social Machines, and to describe the activities of the people and the machines working with their infrastructure. Can we also identify other Scholarly Primitives in our descriptions of Scholarly Social Machines in the previous sections?

In our examples there is typically a stage where content, context and analysis are shared, and responses generated. These seem fundamental and are the basis of the engagement that characterises a Scholarly Social Machine. Hence we might suggest **Sharing** and **Responding** are primitives. These may relate to the Delivering and Collecting primitives in the analysis by Blanke et al [3], who apply Unsworth's Scholarly Primitives method in the context of multiple e-Research projects; they also identify Discovering and Comparing.

We might also expect our primitives to feature some aspect of engaging people in various roles. By way of comparison, a study by Wiggins [27] discusses a typology of citizen science which describes the people involved in creating a crowdsourcing Social Machine

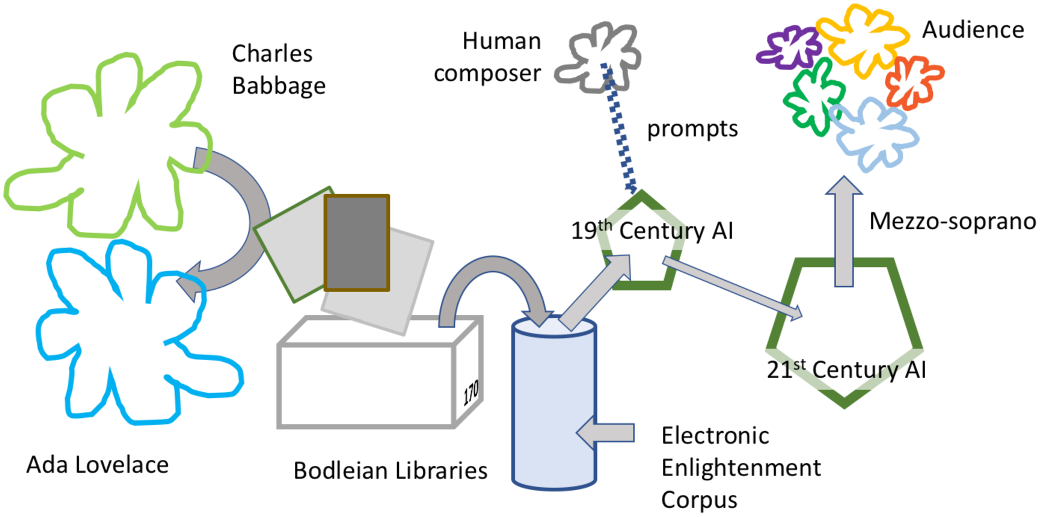

Figure 2: Depiction of the creation of *Alter*

and the nature of their participation; for example

- Contributory—designed by researchers, the public primarily contribute data;
- Collaborative—public help to refine project design, analyse data, or disseminate findings;
- Co-created—some of the public are actively involved in most or all steps of the scientific process.

We refine these two points with reflections based on our analysis:

(1) **The nature of truth.** At first glance, classification of images in Galaxy Zoo is essentially a statistical process: each image is seen by multiple volunteers who make independent classifications. However, the volunteers can optionally engage through forums in discussions with other volunteers and the researchers whose questions motivated the projects, so there is some sharing of practice and exchanging of knowledge, and this process has led to new significant results and discoveries. Contrast this with Wikipedia, where all users can see all edits, and hence the knowledge is explicitly cumulative. These appear to be quite different processes and hence there may be two forms of collecting annotation: **independent** and **cumulative**.

(2) **Degree of co-creation.** How much play is there in the machine? While scholarly communications seem rigid, there are many attempts at interventions, including means of fitting new artefacts into the existing infrastructure through data citation and software citation. Wikipedia provides a remarkably open and accessible platform, which includes for example the contribution of bots for automation. myExperiment was designed for perpetual beta and early usage informed the creation of Research Objects. Zooniverse has evolved to enable anyone to create a project, and project leaders have been open to surprises from volunteers, to the benefit of their research. We might assert that Scholarly Social Machines have room for play. This is not about primitives, but it is about Knowledge Infrastructure.

If we look from the viewpoint of an individual scholar and their workflows, we see cogs in the wheels of the Knowledge Infrastructure, each performing Scholarly Primitives. But what we are observing here is that scholarship now involves the creation of Scholarly Social Machines, and as such we are all creators of Knowledge Infrastructure. Today's Scholarly Primitives therefore reflect the stance of creator as well as user, as researchers and citizens are themselves empowered to create Social Machines. These Scholarly Social Primitives include exploring, connecting, subverting, creating, sharing, and responding.

Finally we note the uptake of automation and machine learning. What does this mean for Scholarly Primitives? It has generated discussion about the need to keep the "human in the loop", and it is interesting to note the Galaxy Zoo enhancement where a percentage of classification tasks are given to human volunteers simply as a calibration and not directly to assist the AI. Perhaps this suggests a role for a Primitive such as *Verification*, but we are as likely to see AI verifying human as *vice versa*. In *Alter*, the work is essentially a co-creation by human and multiple AIs, and perhaps we need to acknowledge that our scholarly output will become this too.

These human behaviours, when brought to bear on Knowledge Infrastructure are the energy that brings momentum to the Scholarly Social Machine. The speed and scale are amplified by co-creation, increasingly learning from each other to discover new insights, as imagined by Berners-Lee and Fischetti. We can benefit, all of us, from the technical and knowledge advances on whose shoulders we stand; we can also now go forward, people and machines together, hand in hand.

7 CONCLUSION

In this paper we have explored the notion of Scholarly Social Machines—the Social Machines in our Knowledge Infrastructure. Through examples we have illustrated the application of this lens, and reflection has refined the notion of Scholarly Primitives to lean into the Social.

Our study of Scholarly Social Machines has given insights that might inform broader Social Machines research. Scholarly Social Machines are: co-created, performed and (consequently) mutable; adaptive; cumulative; and dynamically limited by the current ability of their constituent actors to embrace surprise and disorder. They have: the potential to further knowledge; the potential to create more than the sum of their parts; multiple collaborators, human and machine; connections beyond themselves; systems that can embrace scale; and lifespans described by their usefulness.

Our work also suggests that the methods of Web Science, which study the Web as an evolving artefact, are applicable to the study of our evolving Knowledge Infrastructure. We hope this paper may form an encouragement for further research.

ACKNOWLEDGMENTS

This research was supported by *The Theory and Practice of Social Machines (SOCIAM)* and *Fusing Audio and Semantic Technologies (FAST)* funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under awards EP/J017728/2 and EP/L019981/1, and by The Alan Turing Institute. myExperiment was supported by JISC and Microsoft; the SALAMI project was funded under the Digging into Data programme by JISC, NSF, and SSHRC; work on the Bodleian First Folio was supported through the generosity of its many donors; and the RNCM Centre for Practice & Research in Science & Music (PRiSM) is funded by the Research England *Expanding Excellence in England (E3)* fund. We thank our colleagues on all these projects for many insightful conversations, and also the organisers and participants in NUI Galway's 2015 *Digital Material* conference, the *People and Machines: Co-creating with Heritage Collections* workshop at The National Archives, January 2020, and the 2019 *Discovering Collections, Discovering Communities* (DCDC) conference.

REFERENCES

[1] Sean Bechhofer, Iain E. Buchan, David De Roure, Paolo Missier, John D. Ainsworth, Jiten Bhagat, Philip A. Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius T. Michaelides, Stuart Owen, David R. Newman, Shoaib Sufi, and Carole A. Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (2013), 599–611. https://doi.org/10.1016/j.future.2011.08.004

[2] Tim Berners-Lee and Mark Fischetti. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* Harper San Francisco.

[3] T. Blanke and M. Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems* 29, 2 (2 2013), 654–661. https://doi.org/10.1016/j.future.2011.06.006

[4] Bodleian Libraries. 2012. *Sprint for Shakespeare.* Retrieved 16 May 2020 from http://web.archive.org/web/20120901005606/http://shakespeare.bodleian.ox.ac.uk/

[5] Robert Burton. 1624. *The anatomy of melancholy.* http://gateway.proquest.com/openurl?ctx_ver=Z39.88-2003&res_id=xri:eebo&rft_id=xri:eebo:citation:99857399 2nd edition.

[6] David De Roure. 2014. The Future of Scholarly Communications. *Insights* 27, 3 (2014), 233–38. http://doi.org/10.1629/2048-7754.171

[7] David De Roure, Carole A. Goble, and Robert Stevens. 2009. The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Gener. Comput. Syst.* 25, 5 (2009), 561–567. https://doi.org/10.1016/j.future.2008.06.010

[8] David De Roure, James A. Hendler, Diccon James, Terhi Nurmikko-Fuller, Max Van Kleek, and Pip Willcox. 2019. Towards a Cyberphysical Web Science: A Social Machines Perspective on Pokémon GO!. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. ACM, New York, NY, USA, 65–69. https://doi.org/10.1145/3292522.3326043

[9] David De Roure, Clare Hooper, Megan Meredith-Lobay, Kevin Page, Ségolène Tarte, Don Cruickshank, and Catherine De Roure. 2013. Observing Social Machines Part 1: What to Observe?. In *Proceedings of the 22Nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. ACM, New York, NY, USA, 901–904. https://doi.org/10.1145/2487788.2488077

[10] David De Roure, Clare Hooper, Kevin Page, Ségolène Tarte, and Pip Willcox. 2015. Observing Social Machines Part 2: How to Observe?. In *Proceedings of the ACM Web Science Conference* (Oxford, United Kingdom) *(WebSci '15)*. ACM, New York, NY, USA, Article 13, 5 pages. https://doi.org/10.1145/2786451.2786475

[11] David De Roure, Kevin R. Page, Benjamin Fields, Tim Crawford, J. Stephen Downie, and Ichiro Fujinaga. 2011. An e-Research approach to Web-scale music analysis. https://doi.org/10.1098/rsta.2011.0171

[12] J. Stephen Downie, Andreas F. Ehmann, Mert Bay, and M. Cameron Jones. 2010. *The Music Information Retrieval Evaluation eXchange: Some Observations and Insights.* Springer Berlin Heidelberg, Berlin, Heidelberg, 93–115. https://doi.org/10.1007/978-3-642-11674-2_5

[13] P. N. Edwards, S. J. Jackson, M. K. Chalmers, G. C. Bowker, C. L. Borgman, D. Ribes, M. Burton, and S. Calvert. 2013. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges.* Technical Report. Ann Arbor: Deep Blue. http://hdl.handle.net/2027.42/97552

[14] Michael Mandiberg. 2015. 7,473 volumes at 700 pages each: meet Print Wikipedia. *Wikimedia blog* (2015). https://blog.wikimedia.org/2015/06/19/meet-print-wikipedia/ Accessed: 2020-02-14.

[15] Eric T. Meyer and Ralph Schroeder. 2015. *Knowledge Machines: Digital Transformations of the Sciences and Humanities.* MIT Press.

[16] Dave Murray-Rust, Ségolènene Tarte, Mark Hartswood, and Owen Green. 2015. On Wayfaring in Social Machines. In *WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web.* ACM, 1143–1148. https://doi.org/10.1145/2740908.2743971

[17] Tim O'Reilly. 2005. What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software. (Sept. 2005). https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

[18] Nigel Shadbolt, Kieron O'Hara, David De Roure, and Wendy Hall. 2019. *The Theory and Practice of Social Machines.* Springer, Springer Nature Switzerland AG 2019.

[19] William Shakespeare. 1623. Mr. William Shakespeare's Comedies, Histories, & Tragedies. Bodleian Arch. G c.7.

[20] William Shakespeare. 2020. *The Bodleian First Folio: A digital facsimile of the First Folio of Shakespeare's plays, Bodleian Arch. G c.7.* Retrieved 16 May, 2020 from https://firstfolio.bodleian.ox.ac.uk/

[21] Ray Siemens, Meagan Timney, Cara Leitch, Corina Koolen, Alex Garnett, INKE with the ETCL, and PKP Research Groups. 2012. Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media*. *Literary and Linguistic Computing* 27, 4 (10 2012), 445–461. https://doi.org/10.1093/llc/fqs013 arXiv:https://academic.oup.com/dsh/article-pdf/27/4/445/2773634/fqs013.pdf

[22] Robert J. Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel (Eds.). ACM, 1049–1054. https://doi.org/10.1145/2567948.2579215

[23] Jordan B L Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. 2011. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011.* 555–560.

[24] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. The Singularity is Not near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (Orlando, Florida) *(WikiSym '09)*. ACM, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/1641309.1641322

[25] Ramine Tinati, Elena Simperl, Markus Luczak-Roesch, Max Van Kleek, and Nigel Shadbolt. 2014. Collective Intelligence in Citizen Science–A Study of Performers and Talkers. *arXiv preprint arXiv:1406.7551* (2014).

[26] John Unsworth. 2000. *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?* Retrieved 16 May 2020 from http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html

[27] Andrea Wiggins and Kevin Crowston. 2011. From Conservation to Crowdsourcing: A Typology of Citizen Science. *44th Hawaii International Conference on System Sciences* (2011), 1–10.

[28] Wikipedia. 2020. *Wikipedia:Bots.* Retrieved 16 May 2020 from https://en.wikipedia.org/wiki/Wikipedia:Bots

[29] Zooniverse. 2020. *Galaxy Zoo - The Science behind the Site.* Retrieved 16 May, 2020 from https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/research