There and Here: Patterns of Content Transclusion in Wikipedia

Mark Anderson
Southampton University
Electronics and Computer Science
Southampton SO17 1BJ, UK
mwra1g13@soton.ac.uk

Leslie Carr
Southampton University
Electronics and Computer Science
Southampton SO17 1BJ, UK
lac@ecs.soton.ac.uk

David E. Millard
Southampton University
Electronics and Computer Science
Southampton SO17 1BJ, UK
dem@ecs.soton.ac.uk

ABSTRACT

As large, collaboratively authored hypertexts such as Wikipedia grow so does the requirement both for organisational principles and methods to provide sustainable consistency and to ease the task of contributing editors. Large numbers of (potential) editors are not necessarily a sufficient bulwark against loss of coherence amongst a corpus of many discrete articles. The longitudinal task of curation may benefit from deliberate curatorial roles and techniques.

A potentially beneficial technique for the development and maintenance of hypertext content at scale is hypertext transclusion, by offering controllable re-use of a canonical source. In considering issues of longitudinal support of web collaborative hypertexts, we investigated the current degree and manner of adoption of transclusion facilities by editors of Wikipedia articles. We sampled 20 million articles from ten discrete language wikis within Wikipedia to analyse behaviour both within and across the individual Wikipedia communities.

We show that Wikipedia (as at February 2016) makes limited, inconsistent of use of transclusion. Use is localised to subject areas, which differ between sampled languages. A limited number of patterns were observed including: Lists from transclusion, Lists of Lists, Episodic Media Listings, Tangles, Articles as Macros, and Self-Transclusion. We find little indication of deliberate structural maintenance of the hypertext.

CCS CONCEPTS

• **Information systems → Wikis**; *Document structure*; • **Human-centered computing → Collaborative content creation**; *Computer supported cooperative work*;

KEYWORDS

Hypertext, Transclusion, Collaboration, Wikis, Wikipedia, Digital Curation

ACM Reference format:
Mark Anderson, Leslie Carr, and David E. Millard. 2017. There and Here: Patterns of Content Transclusion in Wikipedia. In *Proceedings of The 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4-7 July 2017 (HT17),* 10 pages.
DOI: 10.475/123_4

HT17, Prague, Czech Republic

DOI: 10.475/123_4

1 INTRODUCTION

A large public collaborative hypertext gives free access to allow any person both to read its content and to add to, or improve, the hypertext's data and structure. The hypertext may thus contain the work of many authors, spread across discrete pages. Their varying editing skills can pose a challenge for those trying to maintain the overall coherence and accuracy of the hypertext's content as a whole—as opposed to activity revising individual articles or generating new content. In wikis, where focus is on the rendered page, incremental edits can lead to unseen structural issues. For instance, under 50% of 'articles' in the English Wikipedia are actually content articles, the remainder are re-direction stubs (see Table 2).

The same information may need to be repeated within different articles across a large hypertext. If text is copied, potential exists for thematic drift between different articles through subsequent edits by different authors. Ideally, in order to retain coherence of the hypertext over time, what we call *longitudinal coherence*, content duplication needs to be identified and consistency maintained.

Transclusion [17] offers one means of avoiding duplication. Deliberate and considered transclusional re-use of canonical sources throughout the hypertext can potentially assist with maintaining coherence and avoiding divergent copy. For example, by re-using text summarising a subject in articles referring to that subject. Furthermore, transclusion—if identified up as such—also offers the potential to indicate provenance of re-used text.

It therefore follows that the use of transclusion within a large Web hypertext should increase longitudinal coherence, but it is unclear how widely and how effectively these techniques are used in existing examples such as Wikipedia. Wikipedia's MediaWiki software does support transclusion (see Section 3), but Wiki studies appear to ignore the implied linkage created by transclusion. Despite some analysis as to the functional nature of edits made in Wikipedia [5], no study has been made of the nature of editing as relating specifically to transclusional (re-)use of content. Built-in Wikipedia queries ('special' pages[1]) and API methods can give some indication of transclusion use, but the reports are opaque and do not lend themselves to further exploration, especially as to how or why editors implemented their ideas. Thus more focused study of transclusion is needed.

By analysing the occurrence and nature of Wikipedia content transclusion, the study set out to investigate these questions:

- *Does Wikipedia show evidence of deliberate use of transcluded article content?* If transclusion is used in Wikipedia, then at minimum transclusion mark-up should be detected

[1] See: https://en.wikipedia.org/wiki/Special:WhatLinksHere, on all article pages.

in article source code using transclusion, disparity in usage should become apparent, either within discrete per-language wikis, or between different wikis.
- *Does the nature of transclusion vary between discrete areas within per-language wikis, or between different languages?* By categorising the subject area of any transclusion activity, disparity in use of transclusion should become apparent, both within discrete per-language wikis and between different wikis.
- *Does article content show distinct patterns of transclusion?* If common, transclusion link patterns may be identified which aid those maintaining the hypertext.

2 BACKGROUND

Transclusion, as coined by Nelson in his *Literary Machines* [17], referred originally to a single hypermedia source occurring in multiple places "*Transclusion means that part of a document may be in several places—in other documents beside the original—without actually being copied there*" [18, preface footnote][2]. Subsequently, he re-defined transclusion as "*reuse with original context available, through embedded shared instancing*" [19, p32], tying it more closely to ideas expressed in his Xanadu system with its 'transpointing'[3] windows.

Besides giving a canonical source, the inherent transclusion linkage can help establish provenance and copyright. Nelson held that indication of transclusion is a front-end function of the hypertext's reader (renderer) [18, footnote p2/37]. The technique does not preclude changes in transcluded sources, it left to the user to select which version to link: if the system holds past version(s) of the source these may be linked [18, p2/26]. Web transclusion, e.g for image placement, generally draws material directly from its source meaning that the transcluding document will reflect any change to the source, i.e. the current version. Thus a transcluded source can provide a single, up to date, canonical source for re-use in multiple other contexts.

For the general computer user, pure hypertext systems have largely been supplanted by the more versatile—albeit less rich—World Wide Web. With this move the general understanding of transclusion has broadened to a more general sense of content reuse. Glushko[7, p.231] defines transclusion as "*The inclusion, by hypertext, of a resource or part of a resource by another resource*". Missing from this is Nelson's concept of side-by-side, visually linked, display of source and calling contexts.

Currently, the pre-eminent form of transclusion of Web content occurs in the crafting of advertisements or sponsored content for just-in-time insertion (transclusion) into web pages; transcluded content is brokered in the blink of an eye. Besides the Web itself, this same form of transclusion is active in the 'walled gardens' of social networks such as Facebook where both ads and sponsored articles 'of interest' may be transcluded into a users feed.

There has been some discussion of transclusion of Web hypertext: in general [13], using hypertext [20][22], HTML transclusion [24][21][11][16][12] and XML/HTML transclusion[6]. However, transclusion still remains atypical for hypertextual *writing* for the Web. Research interest tends to focus on either the technical implementation or the social aspect of use. Consideration of the writing of hypertext, in a non-fiction context, can fall between these stools.

Halasz's 'Reflections on "Seven Issues"' [8, p.112] noted that the versioning 'issue' was not fully resolved. In a wiki system [14], the default is to render the current edit of the requested page. All past edits can be rendered and by furnishing the UID of the desired edit. However links, including transclusions, are not tied to a target edit; thus rendered content may change if the transcluded source is edited. For a web-based hypertext wiki supporting transclusion this means, in simplest terms, that the rendered article content (the body copy) of a page is able dynamically to include content not present in the article's own source code. Further indication of transclusion, or ability to traverse such implied links is left to individual implementation.

Transclusion, applied appropriately, could help Wikipedia's many editors maintain cohesion. A precept of Wikipedia quality is the 'many eyes' theory [15]—that many people have looked at any given fact. However, Wikipedia's Manual of Style[4] makes no mention of transclusion (or transcluding from Wikidata), effectively blinding the 'many eyes' to the concept.

Halfaker *et al.*[9] find that there is a plateauing in numbers of active editors of Wikipedia, with the suggestion that there may a natural equilibrium in levels of active editors in collaborative wikis.

Wikipedia has a very flat hierarchy of administrators and users although either of those may have extra roles [1]. There is a notion of a 'quality assurance' role but this seems to apply more to anti-vandalism than hypertextual coherence. For Wikipedia editors kudos is most easily acquired, and thus promoted, by concentration on the 'quality'[5] of individual rendered articles. There appears to be no role or similar incentive for editors maintaining the hypertext's structure in support of its longitudinal coherence (and which may be seen as conflicting with per-article focus).

3 TRANSCLUSION MECHANISMS IN WIKIPEDIA

Wikipedia supports transclusion[6] of the simplest form: all or part of one article may be transcluded into another. Transclusional links are one-way: content is transcluded *into* and article. The link is not locked to particular edit. Thus, regardless of the edit version of the transcluding page, the transcluded content is always from the current edit. A confusing factor for the study is that transclusion—of article content—is effected as a *subset* of general templating functionality. Transclusion markup assumes a target in the `Template`[7] namespace but may target other namespaces, as explained below. Thus any analysis needs to separate out content transclusion from more general utility scripting activity.

A wiki page element (article, template, etc.) named `somepage` can be transcluded into another page via use of a series of general

[2] In the same footnote he records that in the book he actually mistakenly used the word 'inclusion' instead of 'transclusion'

[3] See: http://xanadu.com.au/ted/TN/PARALUNE/paraviz.html and [18, p2/34]

[4] See: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[5] See: https://en.wikipedia.org/wiki/Wikipedia:Featured_articles.

[6] Since 30 May 2004. See: https://en.wikipedia.org/wiki/History_of_Wikipedia#Hardware_and_software.

[7] See: https://www.mediawiki.org/wiki/Help:Templates.

mark-up variations based on the target namespace[8]. The following basic variations occur (as documented[9] in Wikipedia):

- {{somepage}}: this transcludes the page somepage from the Template [sic] namespace, i.e. Template is the default namespace for transclusion assessment.
- {{:somepage}}: this transcludes the page somepage from the main namespace, i.e. a content article.

The second form above is therefore the canonical marker for *content* transclusion. Various tag forms based on the above interact to form a confusing and incompletely documented set of alternative transclusion pathways as shown at Figure 1 and described in more detail below.

3.1 Transclusion Mark-up

3.1.1 Tags Marking Transclusion. Incremental improvements to MediaWiki's codebase have resulted in multiple forms of transclusion mark-up. The method termed '*Labeled Section Transclude*' (LST) is a recent addition and allows partial transclusion of specified 'sections' of a called page. A 'section' is either text following a defined heading or ad hoc ranges of text as defined by mark-up in the called page using the <section> tag[10]. For transcluding articles there are five pertinent source code tags, some with aliases[11]:

- {{:somepage}}. Marker: '{{:'. Denotes full article transclusion from somepage unless modified by scope-resisting mark-up within the transcluded article.
- {{:pagename|transcludesection=section_name}}. Marker: '|transcludesection='. Partial transclusion via '<section>' tags or ad hoc section definition in target article.
- {{#lst:pagename|section_name}}. Marker: '#lst'. LST partial transclude. Only section_name is transcluded from pagename. Aliased as #section:.
- {{#lstx:pagename|section_name}}. Marker: '#lstx'. This is an LST exclude. All of pagename is transcluded except section_name. Aliased as #section-x:.
- {{#lsth:pagename|heading_name}}. Marker: '#lsth'. This LST targets headings. Only the content below heading_name up to the next heading of same depth is transcluded from pagename. Aliased as #section-h:.

3.1.2 Tags Controlling Transclusion. Unless an LST or a 'transcludesection' call is used in the transcluding mark-up, it is otherwise not possible to tell whether the article is wholly or partially transcluded. For transcluded articles there are 5 pertinent source code tags:

- <noinclude></noinclude>. Marks ad hoc sections of the called article which are not to be transcluded (but is still rendered for the article itself).
- <onlyinclude></onlyinclude>. Marks ad hoc sections of the called article which are the only parts of an article to transclude (and rendered for the article itself).
- <includeonly></includeonly>. Marks ad hoc sections of the called article which are not rendered *except when transcluded*.
- <section begin="section_name"> <section end="section_name">. Marks ad hoc sections of the called article transcluded by 'transcludesection' calls or via LST.
- <onlyinclude>{{#ifeq:{{{transcludesection|SECTIONNAME}}}|SECTIONNAME| }}</onlyinclude>. Marks an ad hoc section of the called article to be treated, dynamically, as a named transcludable section by 'transcludesection' calls (bullet #2 in list above). The #ifeq method predates LST and the <section> tag.

3.1.3 Visible Indication of Transclusion. All editors, both human and automated, are allowed to use transclusion although it is not self-annotating. It is thus an editor's optional task. English Wikipedia documents allow template-based mark-up to indicate whole[12] or partial[13] transclusion, including a link to the source article.

3.1.4 Transclusion from Wikidata. The 'Phase 2' of the Wikidata[14] project was implemented for all Wikipedia languages in 2013[15], via the {{#property:P_number}} mark-up. However, the method is not yet documented in Wikipedia nor publicly recommended.

4 METHODOLOGY

Our approach deliberately concentrated on transclusion as it affects the task of editing the content of the hypertext rather than the technological aspects of transclusional article rendering. We took the perspective of an editor trying to use transclusion—perhaps for the first time—and having only the information provided *within Wikipedia's documentation* for guidance.

By article *content*, we refer specifically to the whole or partial re-use of copy from *articles*—pages in the main namespace[16] of Wikipedia wikis, i.e. the content as presented via Wikipedia's webpages. Although MediaWiki's transclusional method allows content to be drawn from any namespace in the current wiki, our data analysis focused only on use within the 'main' namespace of each single language wiki within Wikipedia; links to other namespaces were tabulated but not further analysed (except for Wikidata transclusion). 'Active' articles are all main namespace articles excluding redirection stubs.

To reflect the information available to a Wikipedia editor trying to employ transclusion, our understanding of Wikipedia's use

[8] Namespaces are described at https://www.mediawiki.org/wiki/Extension_default_namespaces.
[9] See 'Wikipedia:Transclusion': https://en.wikipedia.org/w/index.php?title=Wikipedia:Transclusion&oldid=693549756.
[10] See: https://www.mediawiki.org/wiki/Extension_talk:Labeled_Section_Transclusion. The introduction date is not documented. The <section> tag is a Wikipedia innovation that predates the HTML 5 <section> tag and there is no functional connection between the two same-named tags, although the MediaWiki tag's mandatory attributes make disambiguation easier. Wikipedia's documentation is ambiguous as to whether '#lst' and'|transcludesection=' are or are not (by design intent) full functional equivalents.
[11] LST Aliases were added to make the mark-up's intent less confusing for inexperienced editors. Wikis may optionally localise for their language.
[12] See: https://en.wikipedia.org/wiki/Template:Transcluding_article.
[13] See: https://en.wikipedia.org/wiki/Template:Transcluded_section.
[14] See; https://www.wikidata.org/.
[15] See: https://blog.wikimedia.de/2013/04/24/wikidata-all-around-the-world/.
[16] i.e. <page> elements in the 'main' XML namespace.

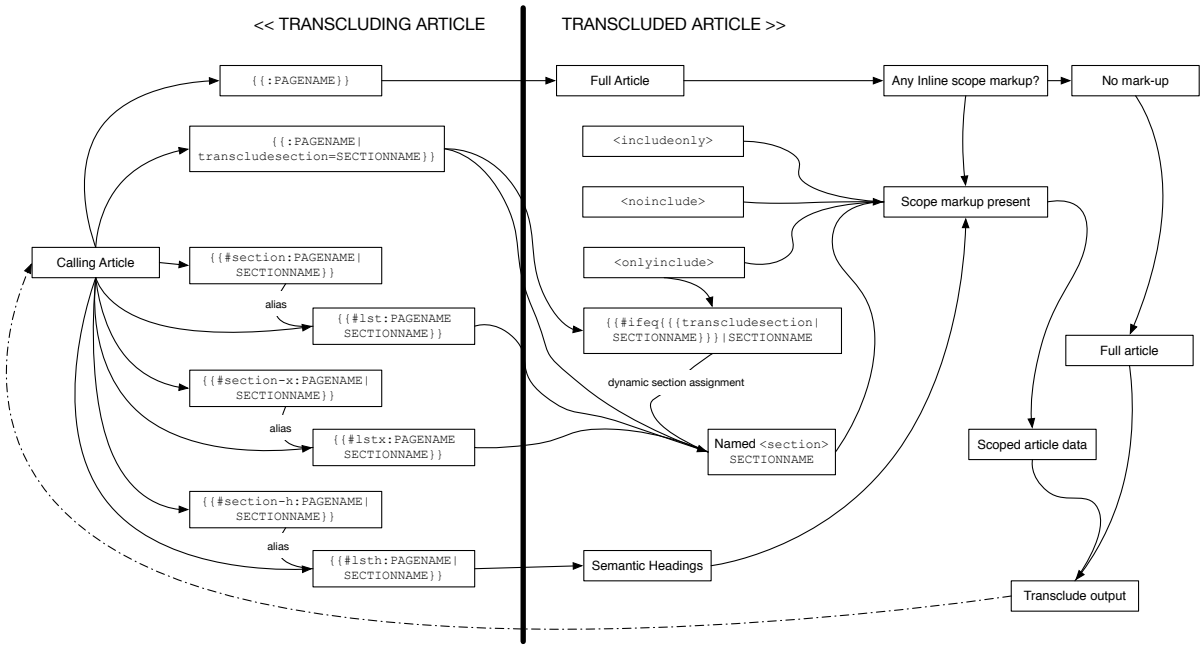

Figure 1: Transclusion pathways in Wikipedia (excluding Wikidata pathway, Section 3.1.4).

of transclusion has been derived from from its own documentation. The wikipedia.org domain serves content in many different languages[17] via per-language sub-domains. For the purposes of this study individual language wikis have been treated as discrete hypertexts.

Working with a static set of data ensured ongoing edits did not affect the transclusion network. Initial sampling test of various static and live Wikipedia data sources also revealed that less ambiguity arose from directly analysing Wikipedia's static XML data dumps than from working with an API into a live wiki. Of the available types of Wikipedia data available for analysis, the monthly 'dumps' of wiki data in XML format were selected as the form of dataset for analysis (see Section 4.1).

To allow exploration of possible differences in transclusion behaviour across differing language communities, data from a number of wikis was collected. Of the languages used in the largest individual wikis, by article count, most are Northern European and use the Roman alphabet. The 10 different languages retrieved are at Table 1; all but Japanese fall within the top 10 wikis by article count Wikipedia. Russian, Japanese and Cebuano were included deliberately, to widen the sample to include wikis using non-Roman characters and non-European languages. English forms the largest wiki in the dataset, containing over 3 times the number of pages of the next largest wiki (Swedish).

URL references to Wikipedia examples have been added as footnotes. Where pertinent, the URLs point to the then-current edit as found in the dataset. In some cases checking details against edits on the live Wikipedia showed problems created by renaming or deletion of articles.

[17] 281 active discrete wikis as at 17 February 2016.

Language	Articles (M)	Pages (M)	Data (GB)	# Files
English (en)	5.06	12.18	56.01	27
Swedish (sv)	2.53	4.14	10.85	1
German (de)	1.90	3.20	10.03	4
Dutch (nl)	1.85	2.52	6.16	4
Cebuano (ceb)	1.81	2.96	7.53	1
French (fr)	1.71	3.11	14.66	4
Russian (ru)	1.28	2.81	16.20	4
Italian (it)	1.25	1.87	9.34	4
Spanish (es)	1.18	2.81	10.24	4
Japanese (ja)	0.99	1.60	9.35	4
Total	**19.57**	**37.22**	**150.37**	**60**

Table 1: Dataset. Wikipedia February 2016 XML dumps. Per-language wiki data, ordered by article count. The Article count includes both live ('active') articles and re-direct stubs. Data date is 13th February for English and 11th February for all other wikis.

4.1 Wikipedia Dataset

XML data from the February 2016 dumps was downloaded (see Table 1) for each of the 10 sampled languages. The dump version chosen was that labelled '*Articles, templates, media/files descriptions and primary meta-pages*', as it proved to have all article source code with a minimum of extraneous material. By comparison, larger dumps included data such as past article revisions and user pages which were not pertinent to the analysis and thus only added overhead to parsing the data.

The chosen dataset provides the content, in XML, of all current articles including re-direction stubs. However, the datasets are undocumented and provide no metadata as to the range and number of namespaces' data included. A complication during initial assessment of data was the lack of any consolidated documentation of namespace titles, including localisations thereof. A primary source used to identify namespace was the (partial) namespace table placed at the head of the XML data files. The XML contains data as `<page>` elements which list namespace, title, and text (source code) of the then-current edit with details of the edit UID and its editor.

4.2 Extraction & Processing

Instance of transclusion markers in source code (see Section 3) were identified via iterative development of a series of Python scripts using regular expressions to detect each of variations of article content transclusion mark-up. English data was used initially and then further localisation added as required, hampered by a lack of documentation of per-wiki namespace and tag localisation strings.

The scripts filtered out all non 'main' namespace XML data elements to interrogate only articles and further filtered active articles from re-direction stubs. The latter are article elements, but can be identified by the presence of an optional `<redirect>` XML element. Of note (Table 2) is the significant number of articles which prove to be simply (hidden) re-directs[18]. The script's namespace filtering design also allowed for re-configured use to look for markers in the Template namespace (the default transclusion namespace).

Of the 10 wiki datasets, 2 have localised[19] aliases defined for the LST hash-based mark-up. Because English versions are supported by default in all wikis regardless of and namespace or tag localisations, this required scripts to test for both mark-up forms.

Initially, extraction scripts generated a Unicode text file per source XML file to assist with resolving detection edge case, helping link detections to the relevant source XML file if extra detail extraction was required. In final form the text files listed the transcluding article's name and each discrete article content transclusion marker within the article; occurrence counts were recorded in a separate file. Items were given additional text delimiters to assist with later separation of articles names for transcluding and transcluded items. For transclusion-limiting tags, discrete start and end tag counts were recorded to indicate whether proper tag closure was being used. Occurrences of 12 discrete transclusional mark-up forms, as described in Section 3, were enumerated for each wiki.

Per-source file output was then aggregated to single file per wiki for each strand of extracted data before data analysis. The generally low occurrences of mark-up allowed analysis using regular expression pattern analysis—and further visual inspection as required. This process enabled identification of edge case detection errors. Besides actual human error in the original article source code, out-of-scope references were tabulated and set aside. Though more laborious, this process gave a richer picture than could be obtained using API query methods.

Further analysis was undertaken in Tinderbox[3], which was chosen for its support for incremental formalisation of emergent structure in the data [23]. Tinderbox data was used to create network data for Gephi analysis of transclusion patterns (as shown in Section 5.3).

5 RESULTS & ANALYSIS

5.1 Occurrence of Transclusion

Evidence of article transclusion was found in every wiki sampled except Cebuano, as shown in Table 2. Despite detection of transclusion, the incidence is very low in comparison to overall article counts. The German wiki showed the highest occurrence rate at 0.58% of all active articles. Aggregating data across all 10 sampled wikis, the averaged transclusion occurrence is 0.095%.

There is no consistency in level of use across the sampled languages. This is also reflected in the occurrence within transcluded articles of the three main mark-up tags used to control the scope of transclusion (as described in Section 5): see Table 3.

The two most-used scoping tags are actually functional equivalents, the German and Russian wikis using a different tag but to similar effect. The German (1.58%) and Russian (1.61%) wikis show similar levels of their most-used tag, although active articles in the Russian wiki represent a sample size 67% that of the German wiki. The German and English wikis favour delimiting source data to be *included* in transclusion whilst the Russian and Italian wikis favour delimiting data to be *excluded*. No annotation was discovered indicating the rationale of these choices, although copying existing practice is a plausible cause.

The amount of articles containing multiple transclusions varied greatly across languages (Table 2). In addition, only 3 wikis were were found to contain articles that both transcluded content and were themselves transcluded: German 614 (5.498% of transcluding articles), English 241 (11.091%), and Russian 34 (2.214%).

The lower occurrence of LST mark-up (see Table 2) must in part reflect its relative newness. Added to MediWikia c.2006, it was cited as unavailable in (English) Wikipedia in mid-2008 and was not added to the Transclusion documentation until early 2014[20] although use of LST has been found as far back as August 2013[21].

The complete absence of content transclusion in the Cebuano wiki, fourth largest by article count, likely reflects the high degree of bot edits. Many of this wiki's articles are stubs created automatically by the activity of bots, such as 'lsjbot'[10]. Absence does not necessarily imply bots cannot program content transclusion as shown by the incidence of Wikidata transclusion (see Table 2). Where found, edits adding Wikidata transclusions are not given explanatory edit comments.

In some cases, use of a particular tag type can be linked to a single editor. For example, all instances of `#lsth` in the German wiki were first added by the same editor (usually without explicit edit comment).

The possibility that a significant amount of transclusion is hidden within templates is discounted. Supplementary sampling of non-main namespaces found `Template` had 7 instances (all English) of content transclusion called from outside the main namespace. The

[18] These arise due to actions like article renaming or (non) use of underscores in URLs.
[19] As described in https://www.mediawiki.org/wiki/Extension:Labeled_Section_Transclusion#Localisation from c.2007. English offers `#section` and German `#Abschnitt` (although the latter doesn't localise `#lsth`).

[20] See: https://en.wikipedia.org/w/index.php?title=Wikipedia:Transclusion&diff=595137158&oldid=586910323
[21] See: https://en.wikipedia.org/w/index.php?title=2013_FIBA_Asia_Championship&diff=567549313&oldid=567548989.

	ceb	de	en	es	fr	it	ja	nl	ru	sv	Total
All 'main' namespace Articles	2,963,362	3,200,021	12,188,486	2,807,709	3,115,027	1,873,355	1,602,047	2,516,924	2,807,922	4,140,672	37,215,525
'Active' (i.e. non re-direct) Articles	1,811,648	1,895,965	5,055,811	1,184,099	1,713,868	1,246,493	998,908	1,850,771	1,281,320	2,533,120	19,572,003
Re-directs as % All Articles	61%	59%	41%	42%	55%	67%	62%	74%	46%	61%	53%
TA Transcluding other Articles.	0	11,167	2,173	70	175	3,046	47	141	1,536	304	18,659
TA as % Active Arts.	0.000%	0.589%	0.043%	0.006%	0.010%	0.244%	0.005%	0.008%	0.120%	0.012%	0.095%
TA Using Multiple Transcludes	0	3,659	933	8	7	2,260	25	103	83	121	7,199
Multiple TA as % TA	0.000%	32.766%	42.936%	11.429%	4.000%	74.196%	53.191%	73.050%	5.404%	39.803%	38.582%
LST Transcludes as % TA	0.000%	0.690%	5.861%	0.000%	0.000%	0.000%	21.277%	0.000%	0.326%	0.658%	1.189%
Wikidata Transclusions as % Active	0.083%	0.000%	0.021%	0.001%	0.087%	0.004%	0.001%	0.020%	0.001%	0.035%	0.026%

Table 2: Per-language occurrence of transcluding articles (TA) in the main namespace.

Language	{{: and LST	onlyinclude	noinclude	includeonly
ceb	0.000%	0.000%	0.000%	0.000%
de	0.589%	1.582%	0.043%	0.008%
en	0.043%	0.225%	0.050%	0.019%
es	0.006%	0.013%	0.116%	0.028%
fr	0.012%	0.007%	0.011%	0.012%
it	0.244%	0.007%	0.595%	0.004%
ja	0.004%	0.022%	0.038%	0.007%
nl	0.008%	0.058%	0.001%	0.000%
ru	0.120%	0.061%	1.615%	0.019%
sv	0.012%	0.046%	0.013%	0.001%
Aggregate	**0.095%**	**0.230%**	**0.172%**	**0.010%**

Table 3: Per-language transclude mark-up occurrence. Column 2: transclusion calls. Columns 3-5: scope-restriction in transcluded pages.

`Portal` namespace contained 29 main namespace transclusions (German 10, English 6, Dutch 1, Russian 12). Nothing was found in the `Module` namespace.

In general, it is hard to assess editors' transclusion intent because where it occurs it is often implemented without any explanation—in either edit comments or talk pages. Unhelpfully, such opaque use informs neither later editors nor a less experienced editor who as yet may not understand the concept of, or rationale for, transclusion. This is reflected in talk page comment from July 2014: "*The word "transclusion," the concepts of transclusion, and code to adeptly accomplish transclusion are not general knowledge. Transclusion is a computer science concept, so little known as to be marked as a spelling error by my dictionary as I work in Wikipedia…*"[22].

Visual marking of transclusion (Section 3.1.3) was not detected, suggesting that the lack of both examples and documentation means editors are unaware of these useful transclusion indicators.

In summary, the fragmented and incomplete documentation, and lack of coherent worked examples obfuscate the transclusion technique for those who might employ it.

5.2 Variation in Purpose of Transclusion

Occurrence counts alone only give part of the picture. Articles in Wikipedia vary in size and scope. Transclusion may simply be more pertinent in some contexts than in others. To investigate this, the transcluding and transcluded articles were reviewed and assigned to broad groupings based on their subject (see Table 4). Transclusions to other namespaces or wikis were assessed as 'out of context'. Articles with code errors or unresolvable transclusion targets were assessed as errors.

For the 9 wikis with transclusions all this involved manual assessment of 20,901 articles; in many cases the title and transclusion targets (translated to English as needed) gave sufficient indication of topic. Where necessary, a smaller number of articles were assessed by direct inspection of the then-current edit. Although the topic choices were subjective, clear groupings did occur perhaps because some topics do indeed lend themselves naturally to transclusion use. For example, sports articles often include team and competition listings. Such tabular data might reasonably be expected to be re-used in multiple contexts, in which context transclusion would aid the process.

Although some groupings were necessarily broad, so as to ensure aggregation of otherwise small discrete topics, the picture that emerged was unexpectedly diverse (see Table 4). The most common topic in aggregate is disambiguation, but it was the most common topic in only 3 of the 9 wikis. However, the use of a pair of 'died on' and 'born on' topics in the Italian wiki might be considered a case of indexing akin to disambiguation.

Whilst each wiki had a predominant transclusion topic, they showed no overall consistency (see Table 4). The most popular topic in each discrete wiki represented over 50% of transclusions, with the exception of Japanese (that also had the fewest total transclusions). The second-most popular topic represents at maximum 26.2% (Dutch), but in most cases is lower, as shown in Table 5.

In the English wiki, the most popular 'Episodic Media' topic covered listings of series and episodes for TV shows or film franchises as well as a smaller amount of printed media such as series of anime magazines and graphic novels. The same topic was seen in only 4 other wikis (see Table 4) and at generally much lower levels.

Despite Wikipedia's categories being a folksonomy and thus not necessarily an accurate listing of article groupings, combining the English wiki's listings of US and of UK TV shows gives 3,974 discrete articles. The intersection with transcluding articles in the Episodic Media topic is only 74 (17%) articles. This indicates transclusion occurrence is low within all possible articles in the overall grouping of Episodic Media. It also illustrates that at least 951 articles relating to episodic media are under-categorised within Wikipedia.

[22] See: https://en.wikipedia.org/wiki/Wikipedia_talk:Transclusion#Please_add_clearer_real_example_text.

Topic:	ceb	de	en	es	fr	it	ja	nl	ru	sv	Aggregate	Agg. %	Rank
Disambiguation		9,820			8				1,284	200	11,312	60.62%	1
Born/Died						2,901					2,901	15.55%	2
Episodic Media			1,125	51			2		129	20	1,327	7.11%	3
Sport		416	251	8	150		17	37		76	955	5.12%	4
Music		602					3				605	3.24%	5
Astromony			243					100			343	1.84%	6
Political			152	4							156	0.84%	7
Computer Games						132					132	0.71%	8
Administrative			85								85	0.46%	9
Film							5				5	0.03%	10
Nature							3				3	0.02%	11
Other Topics <3% ea.		324	271	3	12	4	4	4	99	6	727	3.90%	-
- Other language			5	1	1				3	2	12	0.06%	-
- Other namespace		5	9			9	3		21		47	0.25%	-
- Errors			32	3	4		10				49	0.26%	-
Total	**0**	**11,167**	**2,173**	**70**	**175**	**3,046**	**47**	**141**	**1,536**	**304**	**18,659**	-	-

Table 4: Per-language occurrence of transclusion by subject group, topics ranked by aggregate totals. (Zero % values omitted.)

In the German wiki alone, the Music topic showed a consistent pattern of transclusion—that of a discography into a musician's article. As with the Episodic Media case, the articles which use transclusion are only part of the articles that might do similar. However, creating a separate discography article for artists with an as-yet limited discography, arguably represents limited return on extra work.

Other emergent topics were Sport (in all but Italian and Russian), and Astronomy (English, Dutch). Sporting transclusions show the greatest re-use of tabulated data and listings, and made the greatest use of LST-style transclusion. Sport being a subject likely to have entries in all wikis for some articles and thus useful to compare per-wiki transclusion. Two sports teams' articles were analysed for transclusion and template use: the Boston Bruins ice hockey team ('Bruins') and Manchester United Football Club ('MUFC'). Both subjects had a page in 9 of the sampled wikis (all except Cebuano).

The Bruins articles were found to have a team roster table in 5 of 9 articles, the others being stub pages. Of those 5, only 3 used transclusion and of the 3 only the French[23] wiki transcluded another article[24]: the German and English[25] wikis transcluded a `Template` namespace page. Cross-checking other ice hockey teams' articles, the per-language choice of namespace for defining the team sheet was consistent.

MUFC had articles in all but Cebuano with 4 of the 9 articles had Wikipedia 'featured article' status. No pages directly transcluded content but 6 of 9 used 'navbox' type[26] indexes at page foot. The articles use many tables but, unlike the Ice Hockey pages, these tables are not transcluded. Similarly other UK Premier League club articles (in English) show the same lack of content transclusion. Though visual style is similar, use of (table) templates shows some variation.

In summary, as with general transclusion occurrence, a clear feature of the result was a *lack* of consistent practice. Small groupings of similar type suggest either copying of source mark-up or off-wiki discussion by editors. Edit comments annotating use of transclusion, or intent thereof, were conspicuously absent.

5.3 Transclusional Patterns

Link patterns in various types of hypertext have been documented in the past (Bernstein [2][4]). This task tested if such repeating patterns may be discerned in transclusion linkage within Wikipedia. Due to the volume of data for manual review, only data from the English wiki was fully mapped to identify patterns (9,774 articles). The full tracing of the English wiki also showed that of 2,127 transcluding articles 2,119 (99.624%) used a partial transclude of some form. 241 articles both transcluded other articles and were themselves transcluded. A number of distinct patterns were seen. However, reflecting the limited overall occurrence of transclusion note that most patterns—though distinct when seen—do not occur in great numbers.

5.3.1 Lists: lists from transclusions. These articles create lists, with some or all items being created from transcluded material. Each item is normally a section of the called articles, the degree of transclusion constrained by mark-up. This first pattern is most prevalent in the German and Italian wikis, but found in most languages. In the German (Disambiguation) and Italian (Births/Deaths lists[27]) wikis the main topics make greater use of the mixing inline data with transcluded listings. Again, the transcluded articles' own mark-up limits the degree of content re-used in the main listing.

In a few cases, the lists could be large and built entirely from transclusion such as the English wiki's listing of UK diplomatic

[23] See: https://fr.wikipedia.org/w/index.php?title=Bruins_de_Boston&oldid=122040311.
[24] See: https://fr.wikipedia.org/w/index.php?title=Effectif_actuel_des_Bruins_de_Boston&oldid=120808444.
[25] See: https://en.wikipedia.org/w/index.php?title=Boston_Bruins&oldid=699243091.
[26] See: https://en.wikipedia.org/w/index.php?title=Manchester_United_F.C.&oldid=699017841.
[27] See: https://it.wikipedia.org/w/index.php?title=1798&oldid=70628770.

Topic ranking	DE	EN	ES	FR	IT	JA	NL	RU	SV
1	87.938%	51.772%	72.857%	85.714%	95.240%	36.170%	70.922%	83.594%	65.789%
2	5.391%	11.551%	11.429%	4.571%	4.334%	10.638%	26.241%	8.398%	25.000%
3	3.725%	11.183%	5.714%			6.383%			6.579%
4		6.995%				6.383%			
5		3.912%				4.255%			
Other (<3% ea.)	2.901%	12.471%	4.286%	6.857%	0.131%	8.511%	2.837%	6.445%	1.974%
Out of scope	0.045%	0.644%	1.429%	0.571%	0.259%	6.383%	0%	1.563%	0%
Errors	0%	1.473%	4.286%	2.286%	0.000%	21.277%	0%	0%	0%

Table 5: Per-topic contribution to overall transclusion occurrence, by language.

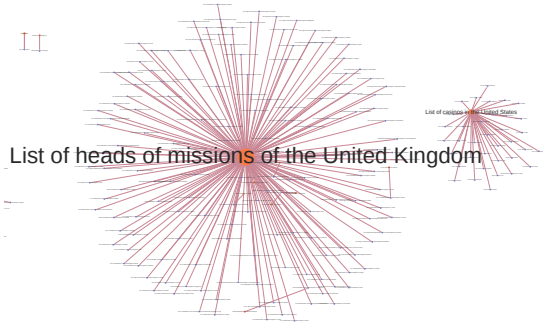

Figure 2: List of UK diplomatic reps: English wiki.

representatives[28] (see Figure 2): this article uses 205 discrete transclusions drawn from 160 discrete sources[29]. The article is also unusual in the care taken to add HTML comments to the source of both calling and called pages to ensure editors understand the process.

The Japanese wiki has a listing of film box office figures[30] using 145 transclusions (of between 1 and 5 instances) of 65 discrete source articles each representing per-year data. This was also the largest transclusion found that used entirely LST mark-up.

5.3.2 Lists: lists of lists. This form is best seen in the listings of minor planets, as found in the English[31] and Dutch[32] wikis. In this pattern for structured data, summaries of lower-level listings are transcluded into a more abstracted listing in the level above. The pattern is easily extended, encompassing large numbers of articles (see Table 4, Astronomy data).

5.3.3 Lists: Episodic Media listings. This is similar to the last but specifically reflects the structure of a show/publication article's listings by season/series summaries and then per-season articles listing episode synopses. In some cases all 3 levels are connected by transclusion (see Figure 3), or else just two; some shows also transclude in character lists. While newer shows may lack content, and older ones verifiable data, there is no clear sign as to why some editors only connect two levels.

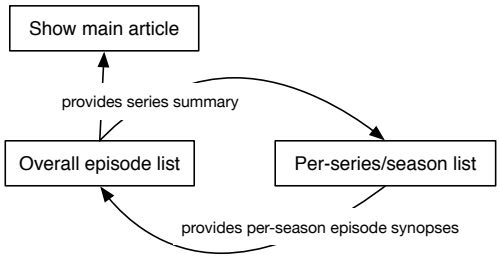

Figure 3: Episodic media listing structure.

Summaries, for upward transclusion of synopses, are consistently defined using scope-limited mark-up. However, choice of scope-limiting tag (Section 3.1.2) varies, with no documented rationale for differing use. It is likely current practice represents initial rote copying of early examples with subsequent divergence through error or further customisation. The English wiki does have some code examples that, if found by a user, can assist with consistent practice but the degree of variation (as seen at source code level) does point to a lack of coherent practice. For instance, the Episodic Media topic appears as many small clusters of 2-3 levels, the cluster size reflecting the number of series (see Figure 4).

There are remarkably few instances of linking between these clusters, one such is shown at Figure 5 (note also a 3-level cluster bottom right). The centre of this largest small tangle is the article "List of longest-running U.S. primetime television series"[33], transcluding articles for shows such a *The Simpsons* and *South Park*.

Cross-cluster linking is most seen in the 'Sport' topic area. Some competitions have many rounds (clusters) forming the route to major championships (Figure 6).

5.3.4 Tangles. If list groupings are interconnected, a 'tangle' can form, with no obvious linear structure[34] . The more closely interconnected the different lists the more complex the result. In Figure

[28] See: https://en.wikipedia.org/w/index.php?title=List_of_heads_of_missions_of_the_United_Kingdom&oldid=690855670.
[29] This is because some people hold multiple posts.
[30] See:https://ja.wikipedia.org/w/index.php?title=%E5%B9%B4%E5%BA%A6%E5%88%A5%E6%98%A0%E7%94%BB%E8%88%88%E8%A1%8C%E6%88%90&oldid=58063614.
[31] See: https://en.wikipedia.org/w/index.php?title=List_of_minor_planets:_86001âĂŞ87000&oldid=656043730.
[32] See: https://nl.wikipedia.org/w/index.php?title=Lijst_van_planetoÃŕden_11001-12000&oldid=38287749.
[33] See: https://en.wikipedia.org/w/index.php?title=List_of_longest-running_U.S._primetime_television_series&oldid=699285724.
[34] This echoes the unstructured 'Tangle' pattern in general hypertext patterns [2, p.24].

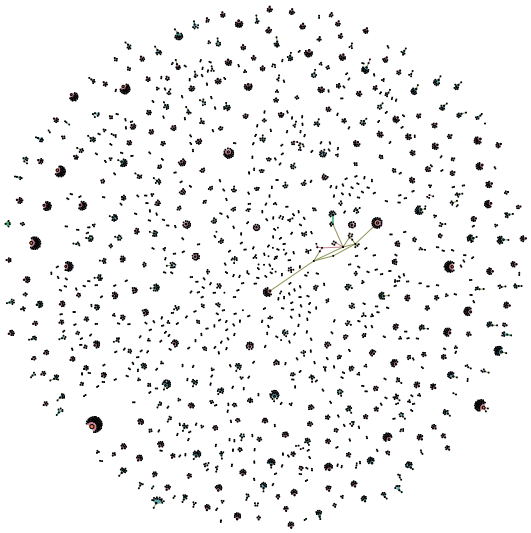

Figure 4: Episodic Media topic transcludes, English wiki. (orange=transcluder, purple=transcluded, green=both)

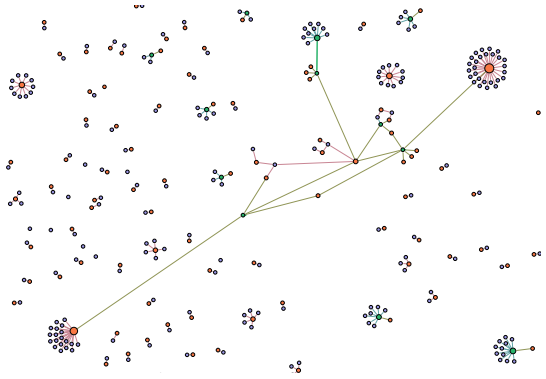

Figure 5: Detail of listing of episodic media (from Fig. 4).

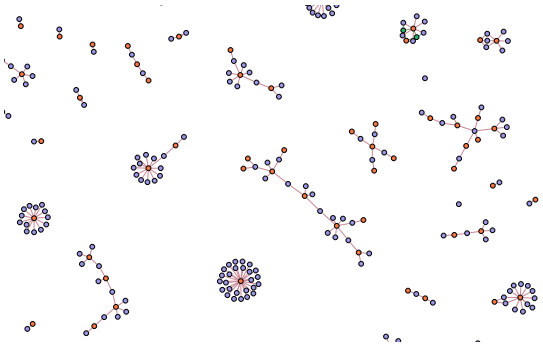

Figure 6: Inter-subject linking, (part of) 'Sport' topic, English wiki.

7, eight discrete listings of villages, towns and cities in Canada, and some per-State listings of communities are co-mingled.

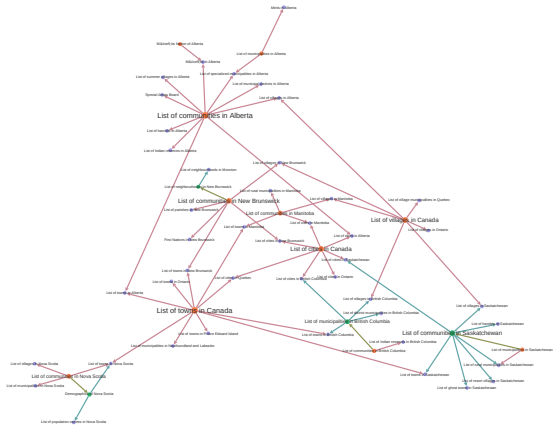

Figure 7: Canadian communities, English wiki.

5.3.5 Articles as macros. An unusual form, seen in Episodic Media in the English Wiki, whereby scope control mark-up excludes all data but the current number of aired episodes. Though the article is transcluded, it functions essentially as a metadata macro. The article on The Simpsons[35] is transcluded for this purpose by 10 other articles. Although this pattern emerged (without edit comment) in October 2015, by February 2016 another 49 Episodic Media articles used the same mechanism to show article count suggesting editors were simply copying a new 'technique'. English wiki 'Portal' namespace articles also transclude counts from 3 different media shows, doing so 5 times in the case of '*The Simpsons*' .

5.3.6 Self-transclusion. An article may transclude itself so as to re-use internal structural content such as list headings. For example, a page on Mediatek tablet processors[36] LST-transcludes a <section>-defined heading to the first table in the page into all the other 15 table headings on the page.

In summary some transclusion patterns were detected. If there were a means to easily visualise these patterns, if could assist with identifying consistent use of transclusion is suitable subject areas.

6 CONCLUSIONS

The analysis undertaken shows that content transclusion definitely occurs in Wikipedia, though at very low levels. Moreover, in subject areas where transclusion might reasonably be applied, it still occurs inconsistently. The current lack of clear documentation, examples and tools to identify transclusion use should be considered a contributory factor. For example, there is no detected use of available templates specifically designed for visibly indicating transclusion (see Section 3.1.3): these assist editors to discern the transcluded material and edit it at source. In addition, although support for transclusion of data from Wikidata was added in 2013 it has yet (as of February 2016) to be added to Wikipedia documentation of transclusion, or to style guides.

[35] See: https://en.wikipedia.org/w/index.php?title=The_Simpsons&oldid=699225461.
[36] See: https://en.wikipedia.org/w/index.php?title=List_of_devices_using_Mediatek_tablet_processors&oldid=691189367.

Tools available to editors to detect transclusion, if not explicitly marked by the originator, are also limited[37] making it difficult even for a diligent editor to detect easily any transclusions created by other editors (especially if no edit comment is used).

Transclusion is found in Wikipedia but appears to be used in a fragmented manner. Whilst MediaWiki software allows for quite nuanced use of transclusion, the full range of capability is not employed. Editors would be better served if transclusion was referenced explicitly in existing style and writing guides. These should explain what sort of topics do or do not suit transclusion and why, along with worked examples to explain the necessary mark-up and the sort of signposting that should be left for other editors. If it were possible to add an option to version-link transclusions (or have a method to record the erstwhile edit UIDs of calling and called pages) this would also benefit long-term maintenance.

With better identification of areas—or categories—suitable for transclusion there is also scope for bots to check on transclusion presence and indicate article clusters that might be improved by transclusion. Such additional structure would also be of use to Semantic Web and data interfaces to the hypertext.

In highlighting apparent weaknesses in documentation, this is not meant to disparage the general efforts of Wikipedia's editors. Rather, it points to a lack of a support role that considers the hypertext as a whole. Transclusion patterns can help both as usage examples and as markers which hypertext maintainers may use to looks for subject areas where transclusion is ineffectively used. The existing roles available to Wikipedia editors do not, as yet, have a niche for those focussed on maintaining the hypertext as a whole. Having a more clearly defined role of this type would make it easier to identify and mediate conflicts between editors (and automated bots) operating in narrow versus broad scope. The page 'Wikipedia:Transclusion costs and benefits'[38] states: "*There is a social cost of transclusion, the total expectation over time of the risk that a transcluded template page may be vandalized*.". This indicates a possible misalignment of interests between article-centric editors and those looking to maintain the lager corpus of work.

WikiProjects[39] is an initiative which focus volunteers on specific subject areas within Wikipedia. Recent analysis of WikiProjects found that "*WikiProjects has reconfigured the article production and improvement process*" and this suggests they may be a possible catalyst area to nurture some more deliberate cross-wiki structural maintenance for the long term. For future work, it is intended to conduct interviews with WikiProject volunteers to assess if their editing intent is mainly article-focused or takes a wider view such as might aid longitudinal curation of Wikipedia.

As the the web matures the effort to support large hypertexts such as Wikipedia will shift from growth to maintenance - with an emphasis on what we have called longitudinal coherence. The infrequent and inconsistent use of transclusions in Wikipedia indicates that approaches that could make this maintenance more manageable have yet to be embraced, although the presence of particular patterns of transclusion do demonstrate their potential. More broadly, a challenge for collaborative hypertexts is to consider the long term and not just the visible 'front page' of content.

[37] For example,'what links here' queries that list inbound links, redirections and transclusions.

[38] See: https://en.wikipedia.org/wiki/Wikipedia:Transclusion_costs_and_benefits.

[39] See; https://en.wikipedia.org/wiki/Wikipedia:WikiProject.

7 ACKNOWLEDGMENTS

This work was supported by the Web Science Centre for Doctoral Training at the University of Southampton, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/L016117/1.

REFERENCES

[1] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional Roles and Career Paths in Wikipedia. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (2015), 1092–1105.

[2] Mark Bernstein. 1998. Patterns of Hypertext. Proceedings of the ninth ACM conference on Hypertext and hypermedia (1998), 21–29.

[3] Mark Bernstein. 2009. Shadows in The Cave: Hypertext Transformations. *Journal of digital information* 10, 3 (2009), 1–8.

[4] Mark Bernstein. 2011. Can We Talk About Spatial Hypertext? Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia (2011), 103–112.

[5] Philip Boulain, Nigel Shadbolt, and Nicholas Gibbins. 2009. Studies on Editing Patterns in Large-scale Wikis. In *Weaving Services and People on the World Wide Web*. Springer, 325–349.

[6] Angelo Di Iorio and John Lumley. 2009. From XML Inclusions to XML Transclusions. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (2009), 147–156.

[7] Robert J. Glushko. 2013. *The Discipline of Organizing*. MIT Press. 752 pages.

[8] Frank G Halasz. 2001. Reflections on "Seven Issues": Hypertext in the Era of the Web. *ACM Journal of Computer Documentation (JCD)* 25, 3 (2001), 109–114.

[9] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[10] E.E. Jervell. 2014. For This Author, 10,000 Wikipedia Articles Is a Good Day's Work. (2014). Retrieved 20 July, 2016 from http://www.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001

[11] Clemens N. Klokmose, James Eagan, Siemen Baader, Wendy Mackay, and Michel Beaudouin-Lafon. 2016. Webstrates: Demonstrating the Potential of Shareable Dynamic Media. Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (2016), 61–64.

[12] Josef Kolbitsch. 2005. Fine-Grained Transclusions of Multimedia Documents in HTML. *J. UCS* 11, 6 (2005), 926–943.

[13] Harald Krottmaier and Denis Helic. 2002. Issues of transclusions. Proceedings of E-Learn (E-Learn 2002) (2002), 1730–1733.

[14] Bo Leuf and Ward Cunningham. 2001. *The Wiki way Quick Collaboration on the Web*. Addison-Wesley Boston, MA.

[15] Andrew Lih. 2004. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature* 3, 1 (2004).

[16] Hermann Maurer and Josef Kolbitsch. 2006. Transclusions in an HTML-Based Environment. *CIT. Journal of Computing and Information Technology* 14, 2 (2006), 161–173.

[17] Theodor Holm Nelson. 1965. Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate. Proceedings of the 1965 20th National Conference (1965), 84–100.

[18] Theodor Holm Nelson. 1982. *Literary Machines*. Mindful Press.

[19] Theodor Holm Nelson. 1995. The Heart of Connection: Hypermedia Unified by Transclusion. *Commun. ACM* 38, 8 (1995), 31–33.

[20] Hartmut Obendorf. 2004. The Indirect Authoring Paradigm – Bringing Hypertext into the Web. *Journal of Digital Information* 5, 1 (2004).

[21] Andrew Pam. 1997. Fine-Grained Transclusion in the Hypertext Markup Language. *Project Xanadu Memo* 2 (1997).

[22] m. c. schraefel, Leslie Carr, David De Roure, and Wendy Hall. 2004. You've Got Hypertext. *Journal of Digital Information (JoDI)* 5, 1.253 (2004).

[23] Frank M Shipman III and Catherine C Marshall. 1999. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)* 8, 4 (1999), 333–352.

[24] Giuseppe Sindoni. 1999. Incremental Maintenance of Hypertext Views. In *The World Wide Web and Databases*, Paolo Atzeni, , Alberto Mendelzon, , and Giansalvatore Mecca (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 98–117.

[25] Ramine Tinati and Markus Luczak-Roesch. 2017. Wikipedia: A Complex Social Machine. *SIGWEB Newsletter* Winter (2017), 6:1–6:10.