Exploring Relations in Neuroscientific Literature using Augmented Reality: A Design Study

Ivar Troost
Utrecht University
Utrecht, The Netherlands
i.o.troost@uu.nl

Ghazaleh Tanhaei
Utrecht University
Utrecht, The Netherlands
g.tanhaei@uu.nl

Lynda Hardman
CWI
Amsterdam, The Netherlands
Utrecht University
Utrecht, The Netherlands
lynda.hardman@cwi.nl

Wolfgang Hürst
Utrecht University
Utrecht, The Netherlands
huerst@uu.nl

ABSTRACT

To support scientists in maintaining an overview of disciplinary concepts and their interrelations, we investigate whether Augmented Reality can serve as a platform to make automated methods more accessible and integrated into current literature exploration practices. Building on insights from text and immersive analytics, we identify information and design requirements. We embody these in DatAR, a system design and implementation focussed on analysis of co-occurrences in neuroscientific text collections. We conducted a scenario-based video survey with a sample of neuroscientists and other domain experts, focusing on participants' willingness to adopt such an AR system in their regular literature review practices. The AR-tailored epistemic and representational designs of our system were generally perceived as suitable for performing complex analytics. We also discuss several fundamental issues with our chosen 3D visualisations, making steps towards understanding in which ways AR is a suitable medium for high-level conceptual literature exploration.

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; *Walkthrough evaluations*; *Systems and tools for interaction design*; *Visualization systems and tools*; *Virtual reality*; • **Computing methodologies** → *Knowledge representation and reasoning*.

KEYWORDS

immersive analytics, augmented reality, linked data, literature exploration, neuroscience

ACM Reference Format:
Ivar Troost, Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2021. Exploring Relations in Neuroscientific Literature using Augmented Reality: A Design Study. In *Designing Interactive Systems Conference 2021 (DIS '21), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461778.3462053

DIS '21, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8476-6/21/06.
https://doi.org/10.1145/3461778.3462053

As more papers are published, the harder it becomes for scientists to maintain an overview. While literature reviews serve to defragment contributions, manual review is costly and – even when performed rigorously – researchers run the risk of missing important perspectives not identified at some part of the review process [12]. Auer et al. argue that this bottleneck is inherent to *document-centric publishing*, and advocate for a move towards "expressing and representing information as structured, interlinked and semantically rich knowledge graphs" [2, p1].

To further investigate how literature exploration practices can be technologically supported, we took neurosciences as an example research field. Talking with neuroscientists at the Institute of Automation of the Chinese Academy of Sciences, we found that one of the main shortcomings of manual literature exploration lies in performing complex relation exploration (Cunqing Huangfu, personal communication, June 12, 2019; July 3, 2019). Without automated tools, it would be hopeless to find out which brain region is most often referenced when discussing a disorder, e.g., depression – indicating wide consensus on this relationship. Likewise, it would not be possible to find out which brain regions are mentioned only seldom in connection with a specific disorder – which could offer fruitful grounds for further investigation.

On both the computational and visualisation side, such as [9, 11], computer scientists have posed the need for a "distant reading" approach to academic literature, i.e., computer-assisted analysis of larger bodies of text. We are not aware that any such system has yielded significant user adoption, however. More work is required to make tools directly accessible to (neuro)scientists, without the need for a supporting data analyst, and to integrate these into existing research workflows.

At the start of our design process, we investigated whether Augmented Reality (AR) could complement current literature exploration practices – both as performed on a 2D screen and in physical space. Recent work in Immersive Analytics (IA; [21]) highlights the intuitiveness of natural user interfaces, as well as the immersive

properties of IA visualisations and the opportunity to render three-dimensional data binocularly (e.g., brain visualisations). AR also allows visualisations in the virtual world to be placed adjacent to a 2D screen or printed paper, and conducting literature explorations over longer periods of time could benefit from persistently placed virtual objects in physical space (cf. *Method of Loci*). Building on design work in [28], we present the design process and evaluation of a novel analytics system implementation: DatAR. We seek to harness the visuospatial representations that AR enables to provide access to neuroscientific knowledge graphs. Our goal is to support neuroscientists in performing complex relation exploration tasks through human-in-the-loop analytics.

We evaluated our system using a scenario-based video survey[1]. While COVID-19 forced our hand in choosing this type of evaluation (due to the inability to safely perform user tests), it allowed us to focus on the quality of our design choices and participants' willingness to adopt the system in their regular practice. For the video survey, we used a mock AR environment in VR (which could be considered a variant of the virtual field study, [18]). This way we alleviated the immaturity of hand-tracking technology that we experienced in earlier design cycles. Our assumption in doing this research is that the quality of hand-tracking technology (and AR technology in general) will improve over the coming years, which is the timeline in which DatAR will become actually useful in practice (cf. Future Studies methods in HCI, [19]). Given our main interest is the willingness of participants to adopt technology in the future, using a higher-fidelity VR-based AR mock-up made more sense than showing an AR version with contemporary limitations.

1 RELATED WORK

1.1 Literature-Based Discovery

Literature-based discovery uses various statistical and machine-learning techniques "to exploit already known scientific knowledge to generate hitherto unknown but meaningful connections" [11, p2]. Using this information, scholars can make better decisions as to which hypotheses to pursue next. The field started off by using *co-occurrence approaches* [27] based on an insight from linguistics, the *distributional hypothesis*: if two concepts are repeatedly mentioned in the same analysis unit (e.g., in a paper, abstract, or sentence), they are semantically associated.

The graph-based approach taken by Gramatica et al. [14] is comparable to that of the providers of our primary data set (the *Brain Association Graph*; see the DatAR System Design section below), uses knowledge graph analysis to find high-potential drug–disease combinations that have not yet been studied in-depth. The knowledge graph is constructed by mining concept co-occurrences in PubMed paper abstracts. The concepts were derived from a predetermined dictionary. Subsequently, they used graph algorithms (e.g., random walk distance) to find the shortest paths between drugs and diseases. Based on their results, they determined several new uses for existing drugs. Our system enables this type of association by co-occurrence analysis interactively in AR.

[1]This footage doubles as a walk-through of the DatAR system, [31] and https://youtu.be/PnOPECRNc_w.

1.2 Topic Model Visualisation

In a regular topic model, an algorithm such as LDA is used to generate n topics based on recurrent word use in a set of given documents. Each document–topic pair is assigned a probability (such that for each document the accumulative probability is 1). This yields an n-dimensional space, which can be visualised by collapsing it to 2–3 dimensions (using a dimension reduction algorithm, e.g., t-SNE). The distance between documents represents their semantic similarity: the closer they are, the more similar (based on the topics assigned). This approach offers an effective means of analysing large data sets through visuospatial distant reading. A good example of this is offered by Li et al. [16], who showcase a multifaceted user interface that allows users to explore subtopics and papers in the computer sciences. The final visualisation offers an effective overview of this document collection, also allowing interactive identification and further inspection of document clusters. Similarly, we use a 3D topic model visualisation to allow users to quickly gauge semantic similarities between neuroscientific concepts.

1.3 Immersive Analytics

Immersive Analytics (IA) attempts to move analytics tools from the 2D screen into our environment (both in Virtual and Augmented Reality; for a survey see [10]). The aim is to design "engaging, embodied analysis tools to support data understanding and decision making [and] liberate these activities from the office desktop" [20, pp14–15].

A key consideration is whether to map data to a 2D or 3D space. Any use of the third dimension has long received strong scepticism within the information visualisation community due to the added visual complexity [20], but there is now a renewed interest in critically re-assessing *binocular* 3D visualisations. As every reduction of an n-dimensional space (such as the output of a topic model) translates to data loss [13], we intend to re-evaluate the merits of 3D information visualisation.

There are several academic toolkits available in the area of IA, such as DXR [26] and IATK [6], based on such works as ImAxes [5]. However, these frameworks were developed with quantitative data in mind. The most important features of the repository for our use case are its graphical structure and associated texts, requiring different visualisation strategies. This is why we developed our system from the ground up.

2 METHODOLOGICAL APPROACH

Our approach is *exaptation*: to take a known – albeit experimental – solution (Immersive Analytics) and apply it to a new problem (literature exploration) [15]. In this early design cycle we report on a situated implementation of artifact [15], with as main purpose to set a course for further theory development in future design iterations. We have taken a constructivist approach, taking notes from Papert's constructionism [23]. We were also informed by distributed cognition, popularised in Information Visualisation [17]. Practically, we are interested in the cognitively coupled system that the user and system represent together, and how coordination between parts of this system allows for successful sense- and decision-making.

We follow the interpretivist framework by Meyer and Dykes [22], who formulated criteria by which a work can be judged after the design process has completed. The authors expect a good design study to be: (1) informed by already existing knowledge, (2) reflexive of the researchers' own role in the study, (3) abundant in having considered and tried many possibilities, and using rich descriptions to convey information, (4) plausible in making knowledge claims that are evidence-based, context-aware and persuasive, (5) resonant by being transferable and evocative, and (6) transparent in being particular enough about reporting. We adopted these values as criteria to meet; a reflection on these can be found in Appendix H of [30].

3 DATAR SYSTEM DESIGN

3.1 Data

The main data set we use is the Brain Association Graph (BAG), developed as part of the Linked Brain Data platform[2]. This is a triple store that contains co-occurrences of neuroscientific concepts in the literature [3]. The repository was created by mining PubMed abstracts for concepts of different types co-occurring within a sentence (for example, hippocampus – a brain region – and Alzheimer's – a disease). Keep in mind that we are not looking at actual medical relationships here – we are looking at how often scholars describe these relationships in their publications. During our experiment, the BAG contained well over 100,000 sentences containing co-occurrences of a brain region and a disease. Statistics were calculated on each concept pair (e.g., hippocampus–Alzheimer's) to determine the total count of co-occurrences, and the two-way probabilities of concepts being in the same sentence. The PubMed ID of the paper of each sentence is kept intact during this process, allowing access to additional metadata (such as date of publication, authors and venue).

While the BAG serves as the central data repository, we (manually) connected its concepts to other linked data repositories to extend the analytic possibilities. MeSH[4] and Wikidata[5] provide additional descriptions of individual brain regions and diseases. We also linked brain regions in the BAG to the Scalable Brain Atlas[6] – allowing volumetric localisation of regions using BAG concepts.

We received a reduced-dimension topic model of brain diseases from Cunqing Huangfu, one of the custodians of the BAG. He used all sentences in the BAG that included at least one region as source data. All sentences describing a disease were combined into single documents, which were then processed with LDA (topic modelling) and T-SNE (dimension reduction) algorithms. This resulted in a three-dimensional coordinate space, in which the distance between diseases represents their semantic similarity (the closer they are, the more similar). This data was used as input to the *Topic Model Visualisation Widget*, described in more detail in the Representational Design section.

3.2 Design Requirements

The following requirements were gathered from the literature, an informal task analysis and several pilot user studies. The provenance and embodiment of these design requirements are further elaborated on in Appendix I of [30].

3.2.1 Supporting Open and Closed Discovery. In literature-based discovery, the distinction is made between open and closed discovery tasks [11]. Whereas closed discovery tasks focus on better understanding the (direct and indirect) relationships between two predetermined concepts, open discovery attempts to identify and study all pertinent relationships originating from a single concept. As our goal is to support *finding* relationships (rather than only inspecting them) our system must support open discovery: Given any concept in the data set, the user must be able to see which other concepts it is related to. Moreover, users should be able to find strong relations (indicating common knowledge) as well as weaker ones (which could offer opportunities for further research).

3.2.2 Highlight, Not Hide. A key feature of relation exploration is that the user does not know what they are looking for prior to their analysis. Therefore, the user has to able to situate themselves in the entirety of the data set. We concur with Woods et al. [32] that *perceptual organisation* that preserves context is therefore a necessity: Users need to be assisted in highlighting subsets of the data to improve observability rather than be faced with a system that hides presumably irrelevant aspects. Highlighting based on perceptual organisation requires modelling the domain semantics (neurosciences), to create a more abstract *conceptual space* to navigate. In our case, we would need to distinguish between the different classes of concepts (regions, diseases, etc.) as well as their co-occurrences (including aggregate statistics). The BAG already offers this structure, challenging us to find suitable ways of organising this data visuospatially such that we utilise high-bandwidth perceptual channels without overwhelming users.

3.2.3 Augment, Not Replace. Our approach should distinguish itself by being fully additional to current (screen- and paper-based) literature exploration practices rather than replacing these. As most such analytics work takes place on the desktop, it would be beneficial to integrate with both physical elements (i.e., Situated Analytics; [29]) as well as implement bridges to other devices.

3.2.4 Making Use of the Medium. AR lends itself to different interactions from traditional desktop environments. Our interface should therefore limit itself to the two affordances dependably supported by hand-tracking technology in AR HMDs: the hand as 3D cursor and grabbing. While keyboard (and other peripheral) support is arguably also part of AR's appeal, as compared to VR, we set as a goal to stick to the core interaction paradigm of AR where possible; any added peripheral would make the system less portable. Instead we focus on *bridging* with other devices and objects, as formulated in the previous design requirement. In addition, we set out to put to use the infinite spatial canvas of AR by building

[2] http://www.linked-brain-data.org
[3] Data is represented in RDF format: triples of Subject-Predicate-Object. E.g., hippocampus-relatedTo-depression and depression-type-disease, where each entity (which can be a concept, a class of concepts, or a predicate type) is represented by a URI. Together these triples form a graph database. For a primer on semantic web notions, see [1].
[4] https://id.nlm.nih.gov/mesh/
[5] https://www.wikidata.org/
[6] https://scalablebrainatlas.incf.org/

a decentralised interface: All functionality should be local rather than global, and there ought not to be any singletons where data views and representations are concerned. This enables an interface "in which multiple queries can be explored simultaneously" [3, p4]. This quality was shown to bolster breadth-first search behaviour in a similar explorative search task in media studies [3].

3.3 Design

3.3.1 Epistemic Design. Based on our review of literature-based discovery (see Related Work) and our design requirements, we support three core tasks for relation exploration in document collections. Firstly, users need to be able to *inspect* lesser known concepts, retrieve their definitions, and find similar concepts. Secondly, they require the capability to *contextualise* a particular concept in regards to a *set* of other concepts. In our system, users can relate any disease to all brain regions, or any region with all diseases. Finally, users should be able to *explain* any given relation found by the system by requesting the sentences that attest to that relationship.

3.3.2 Representational Design. We represent individual concepts (such as "Alzheimer's Disease") as tangible visuospatial entities. This allows users to perform further operations on them to satisfy the epistemic design requirements. We call these representations **Resource Spheres (RSs)**, depicted in Figure 1 (0). RSs contain a Uniform Resource Identifier (URI), a user-facing label, and a class[7]. Users can move RSs using a grabbing gesture, fitting in with the AR interaction paradigm. RSs are used as input to Widgets (or can be output by them). **Widgets** are analysis tools that perform actions such as querying, data manipulation, visualisation or data export.

A Widget can be standalone, requiring no further user input. The *Available Classes Querier* (2), for example, renders all available concept classes in the data storage (e.g., Disease, Region) as RSs without needing user instructions. Other Widgets have one or more **Receptacles**, in which a RS needs to be placed. For example, the *Resource Sphere Inspector* (8) is a Widget that, after placing a RS in its Receptacle, pulls descriptions and closely matched concepts from Wikidata, MeSH and the Scaleable Brain project. Descriptions are then displayed to the user; concepts are output as new RSs. This Widget addresses the Inspect task goal we set. Likewise, the *Concepts of Class Querier* (1) pulls in all Concepts that belong to the provided class.

Some Widgets allow or require a **Dataflow** as input rather than a RS. Dataflows contain a set of concepts, and allow Widgets to communicate with each other[8]. Widgets can have a Dataflow **Outlet** (for Query Widgets), **Inlet** (for Visualisation and Export Widgets), or both (for Manipulation Widgets). An Outlet and Inlet can be connected by holding them together (where each Outlet can connect to multiple Inlets). An example of a Widget using Dataflows is the *Dataflow Inspector* (7), which renders incoming contents as a list. Dataflows allow chaining multiple Widgets, which automatically update if anything prior in the chain is modified.

Our system takes into account the open-world assumption (i.e., our data set could contain concepts of any type, including unknown ones). Widgets are therefore responsible for specifying which data types are acceptable for it to process. For example, the *Co-Occurrence Querier* (3) requires a concept of any type, and a class. It will subsequently try to find relations between the given concept and any items of the given class, and output these as a dataflow. The *Concept Pair Exporter* (9) requires one concept of type Region and one of type Disease, and sends this information to a web-based companion application. This web interface subsequently queries the BAG and PubMed and displays sentences (and their papers) containing both concepts – satisfying the Explain task goal we set and aligning with our "Augment, Not Replace" principle. The *Min-Max Filter* (4) supports Dataflows that pass a co-occurrence list; it then outputs a modified co-occurrence list based on the filter parameters.

A final core mechanic is highlighting. Each item in a Dataflow has a highlight flag, which can be read and/or modified by Widgets. There are two filter states: out of filter range (in red), and (2) in filter range (in yellow). The default/inactive colour is turquoise. This information allows downstream visualisations to render their contents differentially.

3.3.3 Visualisation Design. In this study, we implemented two main visualisations: a Topic Model and a Brain Region Visualisation. Conceptually, the *Topic Model Visualisation* (5) takes in a class (RS) and returns a reduced topic model as a three-dimensional scatterplot of all concepts of that class. Each point is represented as a RS, which is replicated when a user tries to grab it. The *Brain Region Visualisation* (6) behaves similarly, but instead displays the central points of brain regions in the Scalable Brain repository.

Both visualisations become increasingly useful when paired with other Widgets. When using Dataflows to connect the *Co-occurrences Querier* (3), concepts in both sets are highlighted. Adding a *Min-Max Filter* (4) in-between additionally allows differentiating concepts in and out of filter boundaries. This three-Widget set-up is currently the most powerful way of looking at the data within our system, satisfying the Contextualise task goal we set.

As with any complex design, we could never report all design decisions in a paper. For example, what the position of the user should be in relation to a 3D scatterplot. We have documented such considerations in our design and research documentation, keeping a transparent and informed trace of decisions following the design study criteria for rigour by Meyer and Dykes [22][9].

3.4 Implementation

We used Unity (v2019.3.9f1, https://unity.com/) to build our main interface. Dataflows and Receptacles were developed using a reactive framework, UniRx (v7.1.0, https://github.com/neuecc/UniRx), which allows for live-processing of data changes. The AR version of the system was built using the Meta SDK (v2.7.0.38) to support the Meta 2 HMD, optionally using Leap Motion for hand-tracking. The VR version of the system was built on SteamVR (v2.5, SDK v1.8.19, https://github.com/ValveSoftware/steamvr_unity_plugin).

The companion web application was developed in Angular (v8.2.14, https://angular.io/) to allow the main environment to send text-heavy content to a screen for easier reading than AR currently allows. To coordinate this communication, we used a RabbitMQ (v3.8.2, https://rabbitmq.com/) instance as a message broker. The

[7] For example, `lbd:amygdala` as URI, `Amygdala` as label, and `lbd:region` as class.
[8] Dataflows as visual representations of data have been used in VR before, with promising results where it concerns ease-of-use [8].
[9] Detailed design documentation is available on request; a summary can be found in Appendix D of [30].

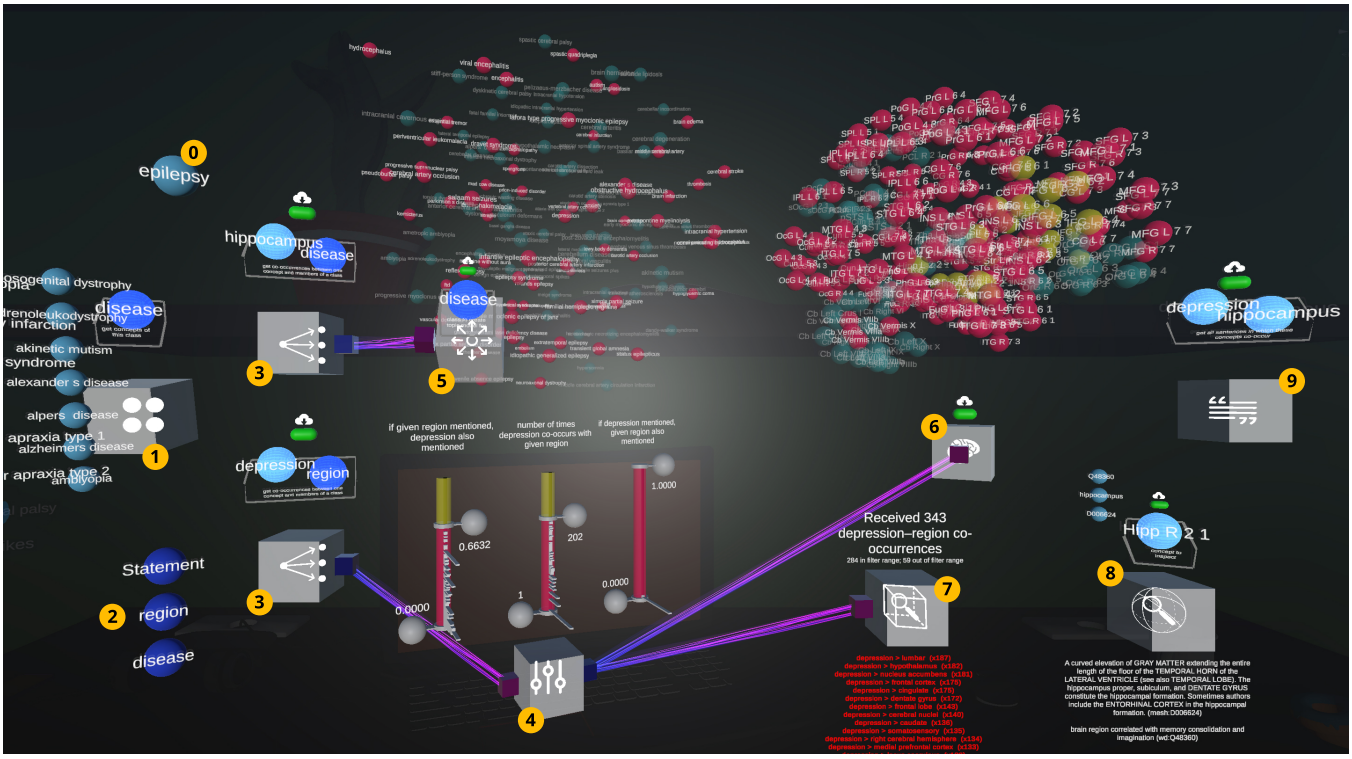

Figure 1: Visual overview of representations in DatAR. The numbered objects are further described in the Representational Design section.

internal data structure in both the Unity and Web interfaces follows the guidelines of the JSON-LD standard (https://json-ld.org/), and is easily exported as such.

Because of the modular implementation of Widgets, contributors are able to easily create new Widgets by combining existing building blocks (such as Receptacles, Inlets, and Outlets), and then adding custom data processing behaviour.[10]

4 EVALUATION STUDY

4.1 Objectives

While we emphasised Perceived Ease of Use (PEoU) and Perceived Usefulness (PU) in our prior pilot user studies, it is useful to look ahead as to whether researchers are waiting for an AR-based literature exploration system as a serious tool in their workflow before too many resources have been expended. What are perceived opportunities in performing relation exploration in AR, and which issues are critical? That is why in this reported design cycle we zoom out and focus on Attitude Toward Using (AT) and Behavioural Intention to Use (BI) systems like ours among a larger and more diverse group of potential users. These are constructs from the Technology Acceptance Model (TAM), a theory from Information Systems that looks at which factors influence adoption of novel technologies [4, 7, 24]. We also evaluate whether participants experience the design artefacts as we had hypothesised in the design rationale, and to what extent our epistemic, representation, and visualisation design are intuitive to grasp.

To these ends, we conducted a video survey. While the disadvantage of this evaluation approach is that users do not get a hands-on experience with the system, it is more accessible and therefore allows a larger and more diverse group to participate. Given our objective and the TAM's focus on *perceptions* of technology, a high-fidelity mock-up scenario is a fitting alternative to user studies.

4.2 Method

4.2.1 Participants. 22 people participated in our video survey (10 women and 12 men, between the ages of 18–64, with a median age range of 25–34). We used a convenience sample, drawing from colleagues of our institutes, a Facebook group of University College students, and participants who had joined an earlier cycle's user study. Seven of our participants were neuroscientists, working in the sub-fields of (cognitive) neurosciences, neurobiology, pharmacology, and psychiatry. Seven participants worked in computer science, data science/visualisation and statistics. Others were active in social sciences (4, of which 3 in educational sciences), digital humanities (1), and military science (1)[11]. Among our sample were participants of various statures, from bachelor's students to full

[10] Access to the source code and a contributor's guide can be provided on request.

[11] Two participants did not fill in a discipline.

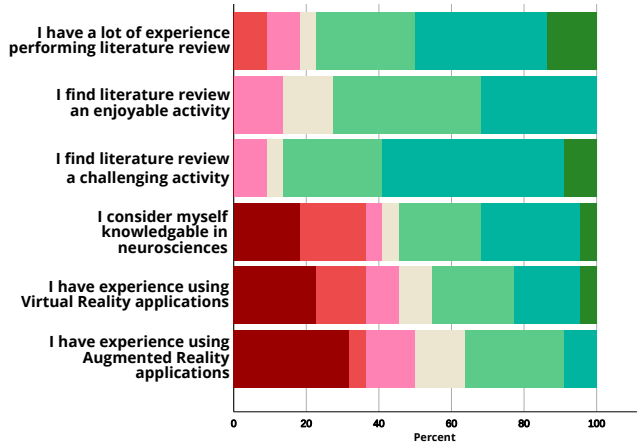

Figure 2: Additional demographics data of our sample (Likert scale from Strongly Disagree to Strongly Agree).

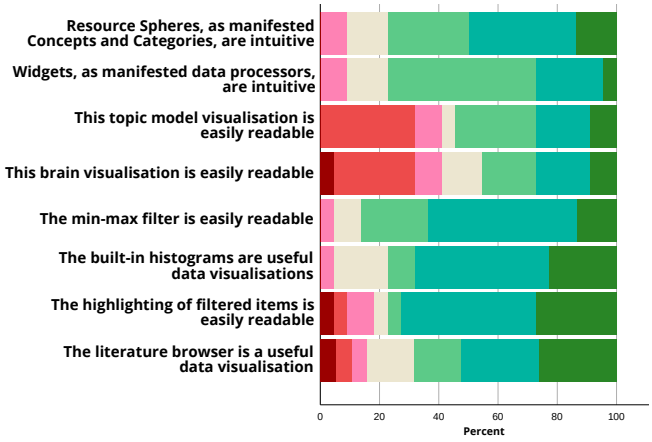

Figure 3: Items posed during video tutorials (section 3; Likert scale from Strongly Disagree to Strongly Agree)

professors[12]. We asked participants about their prior experiences with reviewing literature, the neuroscientific domain, and XR[13] platforms (see Figure 2). We report on correlations with these background factors in the Results section.

4.2.2 Protocol. To gather data, we administered a video survey (which took on average 37 minutes, $SD = 14$). We introduced the survey to participants as a means of "exploring a speculative future in which researchers use AR interfaces to support them in understanding complex relationships in their academic literature." The system shown to users used VR to create a mock AR environment. This set-up was explained in the first video, which also showed an earlier AR version of the system.

4.2.3 Survey. The survey consisted of four sections. Firstly, we inquired about the demographics (reported in the Participants section). Secondly, we asked participants to describe any specialised tools they use for performing literature reviews, to increase our understanding of how users currently tackle similar tasks. The third section contained seven tutorial videos and a mock-up scenario; the footage is available via [31]. According to Kyng, mock-up scenarios aim to "simulate future use situations in order to allow end users to experience what it would be like to work with the system under development and thereby to draw on their tacit, non explicit knowledge and experience" (as cited in [25, p6]). In other words, they are a means to elicit a conversation with the user, which we tried to capture in our (asynchronous) survey by offering ample comment space below each video. For each representation introduced, we also asked participants whether they found it readable, intuitive, and/or useful (on a Likert scale). We treat these questions as conversational triggers, and the answers as spontaneous responses. In contrast, the fourth section asks participants to reflect on several important aspects of our system *after* they have seen them used in context during a mock-up user scenario.

This final section additionally contained sixteen bipolar adjective pairs, which aimed to measure the TAM constructs mentioned earlier as well as two antecedents to PU: Perceived Enjoyment (PE) and Perceived Informativeness (PI). We used the same adjectives as Rese et al. [24] in their analysis of AR applications (see Figure 5). These reflect the TAM model reasonably well, with the added benefit of indicating paths of improvement over more traditional item scales. To allow more granular analysis in our small sample, we deviated from Rese et al. [24] by asking users to fill in a bipolar scale rather than selecting a subset of adjectives; in line with the approach by Davis [7].

4.2.4 Data Analysis. We performed an exploratory analysis on the quantitative data using SPSS (v26.0.0.0), and Kendall's τ_b statistic to compare ordinal associations. We analysed the qualitative feedback, given by participants throughout the survey, using an inductive approach in Nvivo (v12.6.0.959). Responses were first open coded; codes were then aggregated where reasonable. Participants were assigned numerical pseudonyms; the nine participants who considered themselves neuroscientists (separate from their indicated discipline) received an additional n designation (i.e., $P1_n$–$P9_n$ and P10-P22). Minor typos in responses were fixed for readability.

4.3 Results

4.3.1 Quantitative Results. Participant reflections on the interface elements are given in Figures 3, 4 and 5. Given the diversity of our sample, we performed an exploratory analysis to determine whether subgroups significantly differed in their responses. Expertise in topic modelling did not impact perception of our topic model visualisation ($\tau_b = -.063, p = .729$); VR experience did not correlate with the likelihood of using our system in VR ($\tau_b = -.031, p = .862$); AR experience did not correlate with the likelihood of using our system in AR ($\tau_b = -.074, p = .734$); and neither VR nor AR experience correlated with the likelihood of using our system on a 2D screen ($\tau_b = -.267, p = .066$; $\tau_b = -.115, p = .516$).

[12]More specifically: two current bachelor's students (who had also joined our second design cycle's evaluation), three participants with undergraduate degrees, five participants with graduate degrees, six current PhD students, one PhD, and five professors (of which three full professors).
[13]XR, or Cross Reality, encompasses both AR and VR.

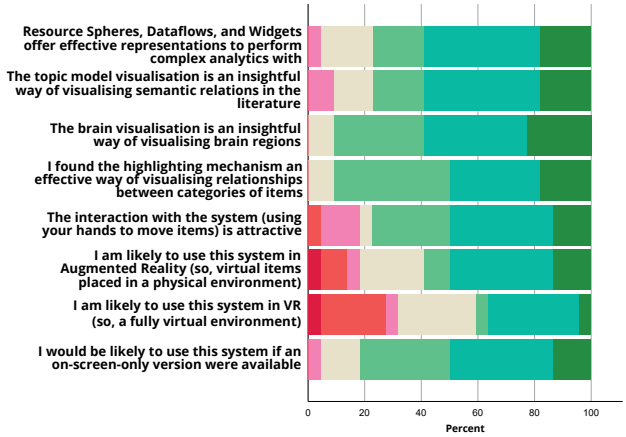

Figure 4: Items posed at the end of the survey (section 4; Likert scale from Strongly Disagree to Strongly Agree)

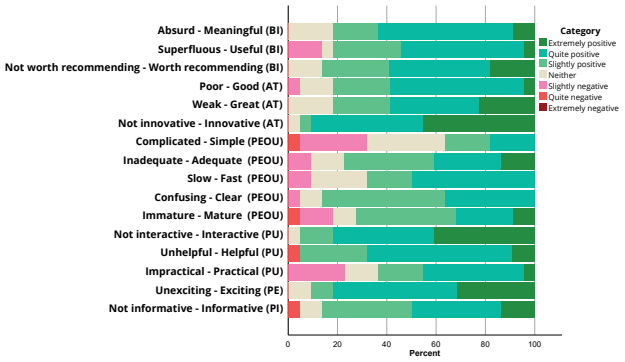

Figure 5: Attitude w.r.t. TAM components: Behavioural Intention to Use (BI), Attitude Toward Use (AT), Perceived Ease of Use (PEoU), Perceived Usefulness (PU), Perceived Enjoyment (PE) and Perceived Informativeness (PI).

Expertise in neurosciences or performing literature reviews caused different perceptions across the 24 non-demographic response items. We found three significant correlations: participants knowledgeable in neurosciences were more likely to use the system in VR ($\tau_b = .338, p = .006$); participants with a higher amount of experience in reviewing literature were more likely to (1) deem the system superfluous (rather than useful; $\tau_b = -.412, p = .002$) and (2) find using hand gestures to control the system attractive ($\tau_b = .323, p = .038$).

4.3.2 *Qualitative Results.*

Literature Review Practices. Thirteen participants (59%) use an *academic search machine*, such as Google Scholar, PubMed, Web of Science, Scopus, PsycINFO, and DBLP. Four (18%) use *Boolean operators* to improve their search queries. For example, $P8_n$ used these *"to capture publications that mentioned several concepts, so that the search is more specific."* Ten participants (45%) use *reference management software* such as Mendeley, Nvivo, Endnote, and Excel. Five participants (23%) use *note-taking and summarisation tools*, such as Microsoft Word, Evernote, and PaperShip. In addition, participants use review papers as a starting point, snowball (repeatedly following references in papers), use data sets, synthesise statistical findings across papers, and use concept-spotting tools (e.g., search-and-highlight in pdf files).

Use Cases. Five participants (23%) framed the system as useful during the early stages of a literature review process. $P9_n$ commented: *"I would probably be interested in viewing data like this for relations I am relatively new to, just to get an idea, but if I am really doing a systematic review I would want to process them one-by-one and explicitly record the themes/dimensions."* $P1_n$ placed the value of the system in use *"during lectures and seminars in the beginning of the bachelors (...). In more specialized regions and higher education the students and teachers have a higher understanding of interspinal relations (...), wherefore this method could be innovative but also time consuming and therefore not as useful as for new students."* P13 commented that the system has *"great potential for complex information seeking tasks, although [they] would prefer traditional systems when [their] tasks are not as complex (e.g., lookup tasks)."* $P9_n$ reflected more critically on where the system could fit in next to existing workflows (also see the Discussion section):

> *My main concern is that I am having trouble imagining the type of review this would be worthwhile for. Most of the time we are researching something that we are already quite familiar with. We have already established an intuition of frontal vs. deep regions, and the visualization may not be so helpful. I feel this is a really helpful way to do a very broad and shallow review, but for me the most laboursome part of the process is retaining this sense of overview while processing the details. After I find a link between depression and the amygdala I would want to know what the papers did. I want to see what the imaging modality was, what the experimental design was, the number of subjects, get a sense of the quality of design/execution, how they analyzed the data, etc. And then after getting that, I would quite like to filter and zoom in and out, but for me, even though this looks exciting, excel would feel a lot easier and less laboursome.*

Finally, $P4_n$ and P10 reflected on the lack of support for *systematic* literature review, wondering *"what the next steps would be in order to write a review and organizing the literature based on these connections"* and posing that *"there needs to be a good way to quantify literature found."*

Fundamental Concerns. While $P4_n$ felt XR visualisations had an advantage over the 2D screen, four other participants were less convinced. P13 and P15 (both computer scientists) want to compare our setup with an equally functional desktop set-up. P15: *"Having the extra burden of VR/AR seems to me to be a distraction, somehow... something that would tire me even more, allowing me to dedicate less time/attention to the actual information being analyzed."* They also pointed out some more fundamental issues with using a 3D topic model: P13 observes that items at its centre are inherently harder to access than those at the periphery; P15 lists several cognitive biases when reading 3D clouds, which need to be mitigated by the

system to guide correct understanding, and wonders whether an abstract document space would be worth this trouble; $P4_n$ worried about decreased legibility if more concepts had to be visualised.

Suggestions for Improvement. A few participants worried about the information density of the two visualisation Widgets, calling them *"overwhelming"* (P12), *"easy to get lost [in]"* (P14), and *"a little vague as there was so much in there"* ($P4_n$). $P2_n$ elaborated: *"I find it very understandable, but not actually readable because all the regions are so close together."* Another concern was the valence and qualitative relationship of co-occurrences ($P4_n$, $P8_n$, P12, & P19). The current system does not distinguish between positive and negative relationships, nor other types of association. To move beyond *"a useful screening tool in an early phase"* (P12), this will need to be supported.

Methodological Comments. Some participants found the voice-over of the video hard to understand, one paused the video to catch some visuals, and one reported nausea due to the VR recording. Comments were given on phrasing and a suggestion to integrate the tutorials in the scenario. More universal were comments either on shortcomings of the (non-interactive) XR video evaluation format, or misconceptions caused by it, for example, it was unclear how to select spheres (by grabbing). Complaints were made about the removal of a RS from the Brain Region Visualisation (which was actually duplicated rather than removed), and about using a rendered laptop screen in VR (which, as explained to participants, was included to simulate an AR environment).

4.4 Discussion

Our results indicate that participants thought that DatAR was innovative, interactive and exciting. A large majority found the representational design easy to grasp and an effective means of performing complex analytics. Participants were, however, concerned about our DatAR's complicated design, and (less so) about its impractical nature and immature state. A major concern is Perceived Ease of Use and conflicting responses were given on the Topic Model and Brain Region Visualisations: their readability was reviewed negatively by almost half of the participants, while the utility of their function was not once reviewed negatively. This suggests that our representational design choices are appropriate but that the visualisation design requires improvement.

Participants offered several suggestions for fine-tuning visualisations, e.g., by using an improved colour scheme. The most important learning point, however, was the need to rethink how to deal with an overwhelming number of data points without losing context (the "Highlight, Not Hide" design requirement). One difficult design decision was how to convey concept names in the Topic Model Visualisation. Our current solution – showing user-facing text labels on all Resource Spheres – was generally perceived as overwhelming. Implementing a metaphorical fish-eye lens may help, emphasising that – while all data is visible – only some is at the centre of attention. Alternative solutions could use visual parameters other than opacity and colour, or alternative filtering concepts.

Platform-level critiques of our system raise a more fundamental question: are the solutions offered by Immersive Analytics a good fit for the activity we are trying to support? More participants rated themselves likely to use the system if it were fully available in a desktop environment rather than in AR (or VR). Likewise, we found that scholars experienced in literature review were more likely to find our system superfluous. However, the same group (to a lesser degree) also deemed hand gesture control more attractive than other participants, suggesting that their notion of superfluousness may refer not to the interaction paradigm, but rather to overall system functionality. This is an important issue to address in future studies. While other factors are at play here (e.g., the AR platform was novel to most users and using a video survey format has some inherent drawbacks in getting across the XR experience), comparative research will be required to make any conclusive statements on whether AR is an appropriate platform for relation exploration in document collections. Particular attention should be paid to critical evaluation of 3D display of abstract data in XR, as user feedback on our current implementation echoes known concerns [20].

Participants noted the system's usefulness in the early stages of literature discovery (for which it was designed), but also indicated that integration with their manual or computer-supported reference management system would be required to support augmentation of current work practices. As stated by $P9_n$, it would also be useful to take an overview of papers in an existing reference management system and visualise and expand the scope using our system.

Some suggested ways to extend DatAR. For example, in education we could facilitate guided explorations of the literature. Others reflected on using the system for disciplines other than neuroscience. While the brain visualisation is specific to neuroscience, all other widgets, data structures and algorithms were designed to be domain-generic. Given similar co-occurrence data, the system could be adapted to visualise other document collections.

5 CONCLUSION

We see a discrepancy between participants' perception of our proposed analytics approach and the use of the XR platform to embody it. This may partly be explained by the use of recordings of a mocked AR scenario rather than having participants experience the system first-hand. However, participants also pointed out more fundamental issues with our chosen 3D visualisations, and the currently problematic integration of shallow relation exploration (in AR) and deep sensemaking (on desktop) – both issues that warrant addressing. Conversely, the AR-tailored epistemic and representational design were generally perceived as suitable for performing complex analytics. We see opportunities in pushing the Resource Sphere–Widget–Dataflow paradigm to its limits by expanding their scope in augmenting intellectual activities.

All-in-all, the reported design cycle and its evaluation study have generated sufficient fuel for a continued exploration of our underlying question: is AR a suitable medium for relation exploration in document collections, and – if so – how? The natural trajectory of this research agenda is towards a comparative study, looking at AR in comparison to the desktop environment given the same task. We must bear in mind that, with any exploration of immature technology, users find it difficult to see beyond their familiar practices to judge the benefits of a complex system with a novel UI at an early stage of development. As designers, we need to identify the underlying tasks of our users, how current solutions fall short in

supporting them and in which respects our proposed solutions can contribute positively. We need to distinguish perceived ease of use issues from fundamental flaws in the epistemic design. Our belief is that, in the complex field of neuroscience, researchers will welcome tools for concept-based exploration of literature. Our prototype AR environment provides a fruitful playground to explore this.

ACKNOWLEDGMENTS

We would like to thank Cunqing Huangfu, Research Center for Brain-inspired Intelligence at the Institute of Automation, Chinese Academy of Sciences, for helping us survey the problem domain and for his attentiveness in resolving any BAG-related data problems. Tessa Heeroma and Anna van Harmelen helped to link various external data sources. The second author is supported by a scholarship from the Iranian Ministry of Science, Research and Technology. This publication is based on the MSc thesis by the first author [30].

REFERENCES

[1] Grigoris Antoniou, Paul Groth, Frank van Harmelen, and Rinke Hoekstra. 2012. *A Semantic Web Primer* (3rd ed.). MIT Press.

[2] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a Knowledge Graph for Science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS '18)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3227609.3227689

[3] Marc M. Bron. 2013. *Exploration and Contextualization through Interaction and Concepts*. PhD Thesis. University of Amsterdam. https://hdl.handle.net/11245/1.399870

[4] Mohammad Chuttur. 2009. Overview of the Technology Acceptance Model: Origins, Developments and Future Directions. *Sprouts: Working Papers on Information Systems* 9, 37 (2009). https://aisel.aisnet.org/sprouts_all/290.

[5] Maxime Cordeil, Andrew Cunningham, Tim Dwyer, Bruce H. Thomas, and Kim Marriott. 2017. ImAxes: Immersive Axes as Embodied Affordances for Interactive Multivariate Data Visualisation. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 71–83. https://doi.org/10.1145/3126594.3126613

[6] Maxime Cordeil, Tim Dwyer, Karsten Klein, Bireswar Laha, Kim Marriott, and Bruce H. Thomas. 2017. Immersive Collaborative Analysis of Network Connectivity: CAVE-Style or Head-Mounted Display? *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 441–450. https://doi.org/10.1109/TVCG.2016.2599107

[7] Fred D. Davis. 1985. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. PhD Thesis. Massachusetts Institute of Technology.

[8] Barrett Ens, Fraser Anderson, Tovi Grossman, Michelle Annett, Pourang Irani, and George Fitzmaurice. 2017. Ivy: Exploring Spatially Situated Visual Programming for Authoring and Understanding Intelligent Environments. In *Proceedings of the 43rd Graphics Interface Conference (GI '17)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 156–162.

[9] Paolo Federico, Florian Heimerl, Steffen Koch, and Silvia Miksch. 2017. A Survey on Visual Approaches for Analyzing Scientific Literature and Patents. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2017), 2179–2198. https://doi.org/10.1109/TVCG.2016.2610422

[10] Adrien Fonnet and Yannick Prié. 2019. Survey of Immersive Analytics. *IEEE Transactions on Visualization and Computer Graphics* (2019). https://doi.org/10.1109/TVCG.2019.2929033

[11] Vishrawas Gopalakrishnan, Kishlay Jha, Wei Jin, and Aidong Zhang. 2019. A Survey on Literature Based Discovery Approaches in Biomedical Domain. *Journal of Biomedical Informatics* 93 (2019), 1–18. https://doi.org/10.1016/j.jbi.2019.103141

[12] Carsten Görg, Hannah Tipney, Karin Verspoor, William A. Baumgartner, K. Bretonnel Cohen, John Stasko, and Lawrence E. Hunter. 2010. Visualization and Language Processing for Supporting Analysis across the Biomedical Literature. In *Knowledge-Based and Intelligent Information and Engineering Systems (Lecture Notes in Computer Science)*, Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer, Berlin, Heidelberg, 420–429. https://doi.org/10.1007/978-3-642-15384-6_45

[13] Antonio Gracia, Santiago González, Víctor Robles, Ernestina Menasalvas, and Tatiana von Landesberger. 2016. New Insights into the Suitability of the Third Dimension for Visualizing Multivariate/Multidimensional Data: A Study Based on Loss of Quality Quantification. *Information Visualization* 15, 1 (2016), 3–30. https://doi.org/10.1177/1473871614556393

[14] Ruggero Gramatica, T. Di Matteo, Stefano Giorgetti, Massimo Barbiani, Dorian Bevec, and Tomaso Aste. 2014. Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action. *PLoS ONE* 9, 1 (2014), 1–10. https://doi.org/10.1371/journal.pone.0084912

[15] Shirley Gregor and Alan R. Hevner. 2013. Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly* 37, 2 (2013), 337–356. https://doi.org/10.25300/MISQ/2013/37.2.01

[16] Zeyu Li, Changhong Zhang, Shichao Jia, and Jiawan Zhang. 2019. Galex: Exploring the Evolution and Intersection of Disciplines. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1182–1192. https://doi.org/10.1109/TVCG.2019.2934667

[17] Hanrui Liu, Chang Liu, and Nicholas J. Belkin. 2019. Investigation of Users' Knowledge Change Process in Learning-Related Search Tasks. *Proceedings of the Association for Information Science and Technology* 56, 1 (2019), 166–175. https://doi.org/10.1002/pra2.63

[18] Ville Mäkelä, Rivu Radiah, Saleh Alsherif, Mohamed Khamis, Chong Xiao, Lisa Borchert, Albrecht Schmidt, and Florian Alt. 2020. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–15. https://doi.org/10.1145/3313831.3376796

[19] Jennifer Mankoff, Jennifer A Rode, and Haakon Faste. 2013. Looking Past Yesterday's Tomorrow: Using Futures Studies Methods to Extend the Research Horizon. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. New York, NY, USA, 1629–1638. https://doi.org/10.1145/2470654.2466216

[20] Kim Marriott, Jian Chen, Marcel Hlawatsch, Takayuki Itoh, Miguel A. Nacenta, Guido Reina, and Wolfgang Stuerzlinger. 2018. Immersive Analytics: Time to Reconsider the Value of 3d for Information Visualisation. In *Immersive Analytics*, Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H. Thomas (Eds.). Springer International Publishing, Cham, 25–55. https://doi.org/10.1007/978-3-030-01388-2_2

[21] Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H. Thomas (Eds.). 2018. *Immersive Analytics*. Springer International Publishing. https://www.springer.com/gp/book/9783030013875.

[22] Miriah Meyer and Jason Dykes. 2019. Criteria for Rigor in Visualization Design Study. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 87–97. https://doi.org/10.1109/TVCG.2019.2934539

[23] Seymour Papert. 1980. *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books, New York, NY.

[24] Alexandra Rese, Daniel Baier, Andreas Geyer-Schulz, and Stefanie Schreiber. 2017. How Augmented Reality Apps Are Accepted by Consumers: A Comparative Analysis Using Scales and Opinions. *Technological Forecasting and Social Change* 124 (Nov. 2017), 306–319. https://doi.org/10.1016/j.techfore.2016.10.010

[25] Antonio Rizzo and Margherita Bacigalupo. 2004. Scenarios: Heuristics for Action. *Proceedings of XII European Conference on Cognitive Ergonomics* (2004), 153–160.

[26] Ronell Sicat, Jiabao Li, Junyoung Choi, Maxime Cordeil, Won-Ki Jeong, Benjamin Bach, and Hanspeter Pfister. 2019. DXR: A Toolkit for Building Immersive Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 715–725. https://doi.org/10.1109/TVCG.2018.2865152

[27] Don R. Swanson. 1986. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine* 30, 1 (1986), 7–18. https://doi.org/10.1353/pbm.1986.0087

[28] Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Huerst. 2019. DatAR: Your Brain, Your Data, on Your Desk - a Research Proposal. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 138–1385. https://doi.org/10.1109/AIVR46125.2019.00029

[29] Bruce H. Thomas, Gregory F. Welch, Pierre Dragicevic, Niklas Elmqvist, Pourang Irani, Yvonne Jansen, Dieter Schmalstieg, Aurélien Tabard, Neven A. M. ElSayed, Ross T. Smith, and Wesley Willett. 2018. Situated Analytics. In *Immersive Analytics*, Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H. Thomas (Eds.). Springer International Publishing, Cham, 185–220. https://doi.org/10.1007/978-3-030-01388-2_7

[30] Ivar Troost. 2020. *Supporting Relation-Finding in Neuroscientific Text Collections Using Augmented Reality: A Design Exploration*. Master Thesis. Utrecht University, Utrecht. https://dspace.library.uu.nl/handle/1874/397219

[31] Ivar Troost, Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2020. DatAR: An Immersive Literature Exploration Environment for Neuroscientists. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 55–56. https://doi.org/10.1109/AIVR50618.2020.00020

[32] David D. Woods, Emily S. Patterson, and Emilie M. Roth. 2002. Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis. *Cognition, Technology & Work* 4, 1 (2002), 22–36. https://doi.org/10.1007/s101110200002