A Standardised Format for Exchanging User Study Instruments

Mark Michael Hall
mark.hall@open.ac.uk
School of Computing and Communications
The Open University
Milton Keynes, United Kingdom

Toine Bogers
toine@hum.aau.dk
Science, Policy and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen, Denmark

ABSTRACT

Increasing re-use in Interactive Information Retrieval (IIR) has been an ongoing aim in IIR for a significant amount of time, however progress has been limited and patchy. While re-use of some study aspects can be difficult due to the varied nature of IIR studies, the use of pre- and post-task self-reported measures is widespread and relatively standardised. Nevertheless, re-use of elements in this area is also limited, in part because systems used to implement them are not able to exchange question, instruments, or complete study setups. To address this, this paper presents a standardised, but extendable, format for IIR survey instrument exchange.

CCS CONCEPTS

• **General and reference → Computing standards, RFCs and guidelines**; *Empirical studies*; *Evaluation*; • **Information systems → Users and interactive retrieval**.

KEYWORDS

IIR, re-use, research design, data exchange, standardisation

ACM Reference Format:
Mark Michael Hall and Toine Bogers. 2020. A Standardised Format for Exchanging User Study Instruments. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343413.3377984

1 INTRODUCTION

As part of the research data management push there has been an increased focus on replicability and re-use of research aspects and components across all disciplines. The widespread use of the Cranfield paradigm to evaluate Information Retrieval (IR) research has meant that IR itself is very advanced in this respect. The widespread sharing of data sets, tasks, and evaluation results via shared activities such as TREC[1], CLEF[2], or FIRE[3] has been a major boon to the field and driven significant progress in the field, in itself validating the value of efforts to support re-use and replicability.

One of the reasons for this is that the system-based focus of traditional IR allows for a high degree of control over all aspects of the experiments, allowing individual parts to be extracted and re-used more easily. The inclusion of the user in Interactive IR (IIR) introduces an element into the research process that is not so easily re-usable and that is tightly coupled to the other study aspects, making re-use of those harder. Attempts have been made at replicating the IR evaluation approach for IIR, however none of these have had long-term success and re-use in IIR remains patchy. However, there are possibilities for improving re-use in IIR, in particular since IIR studies tend to have very similar fundamental experiment methodologies. This is particularly true for the self-reported metrics that are generally acquired from study participants before and after they undertake the study's task(s).

The Barriers to IIR Resource Re-use workshops 2018 and 2019 have been the latest attempts to address this and promote re-use in IIR [2]. One area the 2019 workshop identified as a target was the use of technical means to support the re-use of parts of or complete studies. This paper presents a proposed schema that supports the exchange of the full range of elements used in the self-reported aspects of IIR studies. By simplifying the exchange of these elements, it aims to support and increase the amount of re-use in IIR.

2 RE-USE IN IIR

One of the reasons often given for the low amount of re-use in IIR is the high amount of variation between IIR studies. Study is used flexibly here to cover any experiment that includes human participants and participant-reported responses. This has been studied extensively over the years, ranging from Tague-Sutcliff's 1992 [13] analysis to Yuan and Meadow's examination of measures in 1999 [17] and Kelly and Sugimoto's review in 2013 [9]. The biggest problem this variation introduces, is that it has not allowed for much comparison across series of studies nor for the aggregation of data from multiple studies to test hypotheses across larger data-sets.

A number of attempts have been made to replicate the re-use structure of system-centric IR initiatives. In the TREC Interactive Track (1997–2002) [6], participants used similar protocols, but tested different variables and used partly inconsistent measurements. The INEX Interactive Track (2004–2010) [10] employed standardised tasks and data collection protocols, demonstrating that shared tasks across research institutions are possible in IIR as well. Nevertheless, the overlap between consecutive years was not as high as desired due to changing corpora, IR systems, and participants. The interactive task in the Cultural Heritage in CLEF lab (2013) [15] and the

[1] https://trec.nist.gov/
[2] http://www.clef-initiative.eu/
[3] http://fire.irsi.res.in

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377984

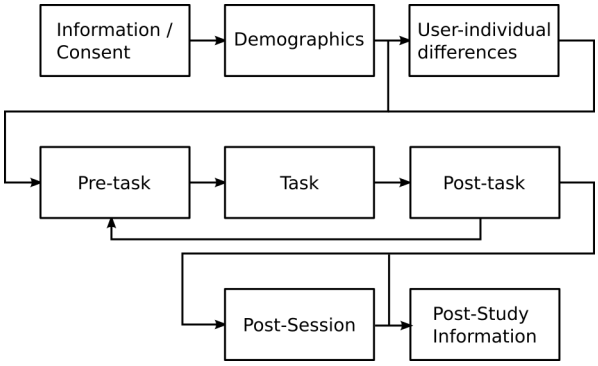

Figure 1: The generic IIR study protocol. As the flow shows, not all steps have to be included for a specific study and the "Pre-task", "Task", and "Post-task" steps can be repeated if necessary.

interactive task in the Social Book Search lab (iSBS, 2014–2016) [4] further standardised protocols and tools used in the shared tasks to increase re-use and comparability of the results. While the final two iSBS iterations achieved a level of standardisation that allowed for merging the results, the task was not continued.

What can be seen through this process, is that it follows the ideas around standardising the IIR study protocol as discussed by Ingwersen [7] and Kelly [8]. This generic IIR study protocol is structured as shown in Figure 1 and is fundamentally linear, with the exception of the possibility to study multiple tasks together with repeating the pre- and post-task steps. Also, the user-individual differences and post-session steps are optional, depending on the study aims. This does not necessarily cover all IIR studies, but due to the flexibility of the specific information acquired in the various steps, it can describe the vast majority of studies.

The generic nature of this protocol has naturally led to the idea that tools can be developed to ease the setup of IIR studies and in the process support re-use of individual parts in the study. The Wi-IRE (Web-based Interactive Information Retrieval) system [14, 16] was one of the first to provide a single system to implement the protocol. A similar system was presented by Bierig and colleagues [1]. While these two systems were designed specifically for IIR, the Experiment Support System [5] is a more generic tool, enabling the researcher to develop a range of complex data-driven studies, ranging from simply questionnaire's to crowdsourcing-style studies and automatically balanced latin-square-based studies. Coagmento v3 [12] is another, very similar system with similar capabilities, but focused more on IIR studies and, as such, better able to support the IIR researcher with the process of setting up their experiment. Additionally there is a range of commercial tools available for building surveys and questionnaires.

While these tools generally allow for re-use of studies or parts of study and both the commercial tools and individual disciplines provide "question banks" to support re-use[4], there is no standard way to re-use elements developed in one tool in another. Our hypothesis is that this is one of the barriers to re-use and this paper proposes a standardised schema to support the exchange of (parts of) studies between systems, thus lowering the barrier to re-use.

[4]For example Psyindex for Psychology Tests: https://www.psyndex.de/

3 USER STUDY EXCHANGE FORMAT

While there is a lot of variation between systems and studies, there is basic commonality, in that studies tend to consist of at least one page containing one or more questions. Depending on the system, there is a varying range of question types available, which the researcher can then customise to create their own questions. Similarly, the transitions between pages can either be linear or, if supported by the system, conditional and include branching and loops. The proposed User Study Exchange Format[5] (USEF) will be able to support this variation, ensuring that it can represent a majority of IIR studies.

The core approach to achieving this is to split all parts of the format into core and optional elements. All systems that can import or export USEF must implement the core elements, but need not implement any of the optional elements. This is the case for the data-types, relationships between data-types, and individual properties on the data-types. The core elements have been kept as minimal as possible in order to lower the bar for USEF adoption as much as possible, while still providing a smooth path to a fuller implementation.

To further ease implementation, the USEF uses JSONAPI as its data-format. In particular it uses compound JSONAPI[6] documents to store all aspects of the study that are to be exchanged in a single file. JSONAPI provides all the required structures, including types, relationships, and properties and due to its use of JSON can easily be processed in any programming language.

3.1 Data Types

When it comes to re-using in IIR, the spectrum goes from re-using individual questions right up to re-using complete studies. In order to support re-use at all these levels, the USEF defines four core types (Figure 2) that cover this range: *Question*, *Page*, *Transition*, and *Study*.

Of these four types, only the *Question* is core and must be implemented, as it is the only type that is guaranteed to appear in all systems. For example a system that provides a hard-coded study

[5]https://github.com/biirrr/usef
[6]https://jsonapi.org/

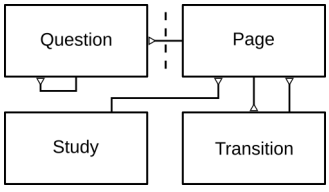

Figure 2: The core type structure of the USEF. Only the *Question* is required, all other types are optional in USEF. *Questions* can derive from another *Question*, to support adaptation and reduce duplication. *Page* and *Transition* are linked via two relationships, as a *Page* has both outgoing and incoming *Transitions*.

and page structures, would not be able to import any of these structures from other systems, however it should be able to ingest and re-use questions from other systems.

A central modelling decision for the core *Question* type was how to distinguish abstract question types, such as a multi-choice question, from specific questions such as an age selection question. While the two could have been modelled as separate types, this would have increased the complexity of the format, required all conforming systems to support both types, and placed limits on re-use. For example, one study's specific question might be re-used with modification in another study, something that is more complex if multiple types are used. Instead the USEF distinguishes abstract and concrete *Question*s. An abstract *Question* is one where the researcher using it must provide some settings before it can be used, while a concrete *Question* is one that could be (re-)used as is. All concrete *Question*s will inherit from one or more abstract *Question*s, but can also inherit from other concrete *Question*s, as detailed below, providing maximum flexibility.

The remaining optional types form a dependency hierarchy, thus *Page*s need to be supported in order to support *Study* and *Transition*s. At the same time, the *Transition* is not linked to the *Study* in order to simplify implementations in systems that only support linear flows.

3.1.1 Question. As stated above, the *Question* represents both abstract and concrete questions. To simplify the creation of both types of questions, the USEF provides a set of core *Question*s that represent the visual elements of the various types of questions that might be used in a study. The focus on the visual elements ensures that the format can be applied to the maximum range of potential questions, as the range of visual elements provided by current user interfaces is much more limited than the range of potential questions. In order to reduce the amount of duplication, when representing studies using in the USEF, the *Question*s implement an inheritance hierarchy. At the root of this is the `USEFQuestion`, which provides the fundamental properties shared by all questions. Based on this, the USEF then provides the eight core *Question*s listed in Table 1.

Each of the abstract types listed in Table 1 is configurable via properties and for each property the USEF supports three possible values:

- null - indicating that this property is not used by this *Question*s,
- {user:xxxx} - indicating that the value for this property is to be acquired from the researcher when setting up a study. Three types of input are defined in the format: single value, multi-line text, and multiple values. Depending on the input type specified, the implementation system needs to provide an appropriate user-input element,
- any other value - indicating a constant value that is to be used for this property.

Properties and their values are inherited from ancestor *Question*s, but can be overridden by descendant *Question*s. This is particularly useful to provided fixed values for previously configurable properties. Figure 3 shows an example of how the "title" property could be overriden in different *Question*s. The root `USEFQuestion` has a "title" property that is null. Its descendant, the `USEFSingleChoice` type overwrites this and specifies that the title is a single value to be provided by the researcher setting up the study. This in turn is overwritten by a custom `FancyGender` type, which provides a fixed, standard phrasing for any *Question* of that type. Similarly, it provides a fixed set of answers and labels for them, to further improve comparability of the resulting participant-acquired data.

Table 1: The nine core abstract *Question*s provided by the USEF. All inherit from the root `USEFQuestion`.

`USEFQuestion`	The root *Question*
`USEFText`	A static text to be shown to the participant. Format of the text can be specified to support HTML, Markdown, ...
`USEFSingleLineInput`	A single-line input element. Specific data-formats (numbers, dates, ...) can be specified via properties.
`USEFMultiLineInput`	A multiple-line text input element.
`USEFSingleChoice`	A single-choice selection from a list of values. How this should be displayed, is configured via properties.
`USEFMultiChoice`	A multiple-choice selection from a list of values.
`USEFSingleChoiceGrid`	A grid containing a series of single-choice selections.
`USEFMultiChoiceGrid`	A grid containing a series of multiple-choice selections.
`USEFHidden`	A hidden input element.

The first two *Question*s in the example are abstract questions, while the third is a concrete *Question* that could be used directly in a study, however both types can be used as parent *Question*s in further *Question*s. For example a `GermanFancyGender` question could extend the `FancyGender` and simply translate all property values into German, to provide a localised version of the *Question*.

3.1.2 Page. A *Page* is an ordered collection of *Question*s, which together form a cohesive unit in the study. These might be a set of questions that together acquire descriptive demographics about participants, but they also might be something more focused, such as a single instrument like the User Engagement Survey [11]. The *Question*s could be both abstract or concrete, allowing for the creation of abstract *Page*s.

3.1.3 Study. The *Study* groups a set of *Page*s into a cohesive whole. The ordering of *Page*s in the *Study* is to be used as the ordering when displaying the *Page*s to participants, except if *Transition*s are also provided, in which case those should be used.

3.1.4 Transition. The *Transition* represents the link between two *Page*s, the source *Page* the participant views first and the target *Page* the participant transitions to after completing the source *Page*'s *Question*s.

A *Page* can have multiple *Transition*s, provided that the *Transition*s have conditions specified on them. In its initial version the USEF only supports a single condition type, namely *Transition*s conditional on responses provided by participants to previous *Question*s. However, future versions are likely to provide conditions based on a list of data-items, to support structures such as looping over a set of tasks, or splitting based on latin-square assignments.

```
{
  "type": "Question",
  "id": "USEFQuestion",
  "properties": {
    "name": "{user:single}",
    "title": null
  }
}
{
  "type": "Question",
  "id": "USEFSingleChoice",
  "properties": {
    "title": "{user:single}",
    "answers": "{user:multiple}",
    "labels": "{user:multiple}"
  },
  "relationships": {
    "parent": {"data": {"type": "Question", "id":
        "USEFQuestion"}}
  }
}
{
  "type": "Question",
  "id": "FancyGender",
  "properties": {
    "title": "Please select the gender you
        identify most with",
    "answers": ["female", "male", "third", "na"],
    "labels": ["Female", "Male", "Third", "Rather
        not say"]
  },
  "relationships": {
    "parent": {"data": {"type": "Question", "id":
        "USEFSingleChoice"}}
  }
}
```

Figure 3: Example definition of three *Questions*, showing inheritance between the three and overwriting of the title property, while the name property is inherited unchanged.

3.2 Extensions & Evolution

The USEF in its core and optional aspects cannot represent all features provided by the various tools that support the researcher in setting up their studies. In order to overcome this, the USEF has an extension mechanism built in, which has adopted from the Web domain. All extensions, whether they are additional types, additional relationships, or additional properties must be prefixed with a string unique to the system that implements them. For example, if a system called "Fancy" implemented *Questions* that were conditional upon the answer to a previous question, then that would add an extra type *FancyQuestionCondition* and add the relationship *fancyConditions* to the *Question* serialisation.

This has two advantages. The main advantage is that it allows for an easier pathway for evolving the USEF. Extension elements that become common across systems can then be integrated into the core or optional parts of USEF, which would essentially simply remove the prefix. Additionally, where similar extensions have been developed by different systems, automated tools can easily be developed to translate between the different versions. This approach has proven itself on the web in the definition of the Cascading Stylesheet format, thus is worth re-using here.

Additionally, all types in the USEF have a version property which allows for continuous evolution of individual parts without having to re-lease new versions of the whole schema. The versioning system will use semantic versioning[7], which allows systems implementing the USEF to automatically detect whether the version of a type that is being imported is compatible with the version implemented in the system. This can then be used to provide the researcher with choices regarding how to proceed when potentially incompatible changes are detected and also to provide tool support for migrating questions.

4 OUTLOOK

One of the barriers in IIR when trying to re-use existing studies or parts thereof, is the difficulty in moving the components to be re-used between different systems set up to deliver the study to participants. In this paper the User Study Exchange Format (USEF) was presented, which aims to provide a standardised format for exactly this exchange. In order to ease adoption, the USEF has a very low entry-bar and then allows for a gradual exchange of more complex study descriptions. It also provides an extension mechanism, to enable study aspects that are not represented in USEF to be included in the USEF representations of a study.

The USEF is in a very early stage of development[8] and input from IIR researchers and tool developers is sought in order to ensure that the USEF is usable in practice and widely adopted. Due to its generic nature, it will also be of interest to researchers in other areas that undertake user studies (psychology, sociology, ...), however initially the focus is on applicability and use in the IIR field.

At the same time, technical availability alone will not ensure widespread uptake. To improve uptake, publication venues (primarily journals and conferences) need to require that for any accepted publication a re-usable representation of the underlying study is also made available. The USEF provides a useful format for this and its inclusion in such a process would set the stage for improved re-use in IIR. To support this, the project will also develop an online tool for manually generating USEF representations of studies, where these have been created in a tool that does not support USEF. A second long-term aim is to also provide a searchable repository for these study representations, similar to RepAST [3] for tasks, which will help with re-use of IIR studies and their elements.

Two big questions for future work are how to evaluate the level of success in supporting standardisation and re-use, and the impact of User eXperience (UX) aspects on re-use. In particular the UX and visual design of a study can have a significant impact on the results and the USEF does not cover these. Whether the USEF needs to be expanded to cover this aspect or whether to treat it as a necessary limitation to ensure adoption requires further consideration.

[7] https://semver.org/
[8] https://github.com/biirrr/USEF

REFERENCES

[1] Ralf Bierig, Jacek Gwizdka, and MJ Cole. 2009. A user-centered experiment and logging framework for interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and interpreting user interactions in information search and retrieval*. 8–11.

[2] Toine Bogers, Samuel Dodson, Luanne Freund, Maria Gäde, Mark Michael Hall, Marijn Koolen, Vivien Petras, Nils Pharo, and Mette Skov. 2019. Overview of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use. In *Barriers to Interactive IR Resources Re-use 2019*. 1–6. http://ceur-ws.org/Vol-2337/overview.pdf

[3] Luanne Freund and Barbara M. Wildemuth. 2014. Documenting and studying the use of assigned search tasks: RepAST. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–4. https://doi.org/10.1002/meet.2014.14505101122

[4] Maria Gäde, Mark Michael Hall, Hugo C. Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Toine Bogers, and David Walsh. 2016. Overview of the SBS 2016 Interactive Track. In *Working Notes of the CLEF 2016 Conference (CEUR Workshop Proceedings)*, Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald (Eds.), Vol. 1609. CEUR-WS.org, 1024–1038.

[5] Mark Michael Hall and Elaine Toms. 2013. Building a Common Framework for IIR Evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 17–28.

[6] William Hersh. 2002. TREC 2002 interactive track report. In *Proceedings of TREC 2002*.

[7] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Vol. 18. Springer.

[8] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1 (2009), 1–224.

[9] D. Kelly and C.R. Sugimoto. 2013. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *JASIST* 64, 4 (2013), 745–770.

[10] Ragnar Nordlie and Nils Pharo. 2012. Seven Years of INEX Interactive Retrieval Experiments – Lessons and Challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peñas, and Giuseppe Santucci (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 13–23.

[11] Heather L. O'Brien, Paul Cairns, and Mark Michael Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28 – 39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[12] Diana Soltani, Matthew Mitsui, and Chirag Shah. 2019. Coagmento v3. 0: Rapid prototyping of web search experiments. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, 367–371.

[13] Jean Tague-Sutcliffe. 1992. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 28, 4 (1992), 467–490.

[14] Elaine G Toms, Luanne Freund, and Cara Li. 2004. WiIRE: the Web interactive information retrieval experimentation system prototype. *Information Processing & Management* 40, 4 (2004), 655–675.

[15] Elaine G. Toms and Mark M. Hall. 2013. The CHiC Interactive Task (CHiCi) at CLEF 2013. In *Working Notes of the CLEF 2013 Conference (CEUR Workshop Proceedings)*, Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro (Eds.), Vol. 1179. CEUR-WS.org.

[16] Elaine G Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra Macnutt. 2008. Task effects on interactive search: The query factor. In *Focused access to XML documents*. Springer, 359–372.

[17] Weijing Yuan and Charles T Meadow. 1999. A study of the use of variables in information retrieval user studies. *Journal of the American Society for Information Science* 50, 2 (1999), 140–150.