

# Portal Design for the Open Data Initiative: A Preliminary Study <sup>†</sup>

Netzahualcoyotl Hernandez \*, Ian McChesney, Joe Rafferty , Chris Nugent ,  
Jonathan Synnott and Shuai Zhang

School of Computing, Ulster University, Newtownabbey BT37 0QB, Northern Ireland, UK;  
ir.mcchesney@ulster.ac.uk (I.M.); j.rafferty@ulster.ac.uk (J.R.); cd.nugent@ulster.ac.uk (C.N.);  
j.synnott@ulster.ac.uk (J.S.); s.zhang@ulster.ac.uk (S.Z.)

\* Correspondence: hernandez\_cruz-n@ulster.ac.uk

† Presented at the 12th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2018), Punta Cana, Dominican Republic, 4–7 December 2018.

Published: 23 October 2018



**Abstract:** The Open Data Initiative (ODI) has been previously proposed to facilitate the sharing of annotated datasets within the pervasive health care research community. This paper outlines the requirements for the ODI portal based on the ontological data model of the ODI and its typical usage scenarios. In the context of an action research framework, the paper outlines the ODI platform, the design of a prototype user interface for the purposes of initial evaluation and its technical review by third-party researchers (n = 3). The main findings from the technical review were found to be the need for a more flexible user interface to reflect the different experimental configurations in the research community, provision for describing dataset usage, and dissemination conditions. The technical review also identified the value of permitting datasets with variable quality, as noisy datasets are useful in the testing of activity recognition algorithms. Revisions to the ODI ontology and platform are proposed based on the findings from this study.

**Keywords:** Open Data Initiative; sharing of annotated datasets; ontology

---

## 1. Introduction

The Open Data Initiative (ODI) has been previously proposed to facilitate the sharing of annotated datasets within the pervasive health care research community [1]. The ODI framework includes a common protocol for data collection, a common format for data exchange, a data repository and related tools to underpin research within the domain of activity recognition. As described in previous work [2], the ODI aspires to use the eXtensible Event Stream (XES) standard for event data storage [3] and an ontological model for the description of experiment metadata.

Work to date has consisted of development of an ontological model and sample instance data as a basis for exploring various usage scenarios [2]. In this paper, we extend this work through the implementation of a front-end portal prototype for the ODI as a vehicle for evaluating the adequacy of the current ODI data model. This is conducted using the action research cycle [4,5] as the guiding research methodology. While more positivistic and quantitative evaluation techniques are envisaged for later iterations of the ODI portal, action research is relevant at this preliminary stage of community engagement and collaboration [6]. It is consistent with the collective contribution of the authors and other technical evaluators to the evolution of the ODI concept and provides a framework for subsequent evaluation stages (Figure 1).

The remainder of the paper is organized as follows: In Section 2 we summarize the functional requirements of the ODI portal. Section 3 introduces the ODI platform and describes the prototype

front-end which is used as a mechanism in this study for evaluating the data requirements, broad usability issues associated with dataset description and document uploading. In Section 4 the technical review is described as part of the “reflect” stage of the action research cycle, including a summary of the main outcomes from this evaluation. The final section presents the proposed revisions and next steps for the ODI project.

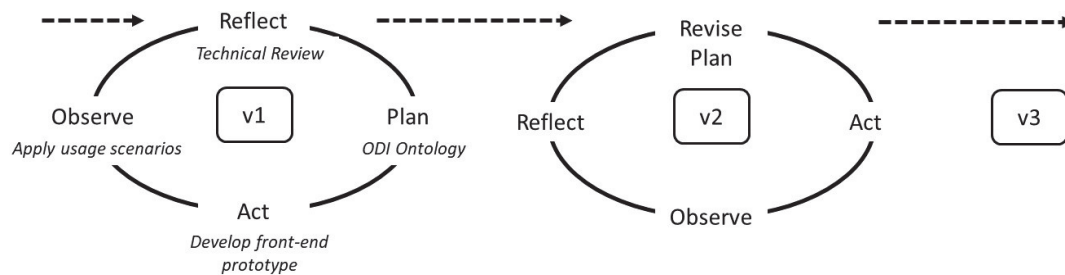


Figure 1. Action research cycle for the ODI portal (based on [7]).

## 2. Requirements for the Open Data Initiative Portal

The ODI portal is a web-based application which allows researchers to register, upload, edit and query experimental datasets. The portal will provide standard registration and login features, enable the upload of experimental datasets, and assist the user in supplying the necessary information to populate the back-end ontology. Where possible, ontology-based reasoning techniques will be used to automatically recognize complex data relationship among the meta-data and datasets submitted to the ODI. Ontology classes and their properties as specified in previous work [2] will be supported in a way which is transparent to the user by virtue of a tiered system architecture. The system shall allow the user to save their work at any time, to preview their upload and to finalize the upload of experimental meta-data and its dataset if and only if all required fields are complete. On completion of the upload function, the system owner of the ODI will be notified of a dataset requiring validation and approval, with user privileges set accordingly for the ongoing editing of unvalidated or validated datasets. A successful upload will be sent to the system data tier. The portal shall also provide a mechanism to support user-defined queries as outlined in the usage scenarios from [2].

## 3. Open Data Initiative—The Platform

The ODI platform is based on a three-tier architecture. The logical viewpoint architecture [8] is shown in Figure 2. The user interface tier enables the collection of information about technology and participants involved in an experimental study, publications and citations and provides the mechanism to upload a new dataset. In addition, it allows a user to construct flexible queries on existing datasets. The business tier includes the application logic, such as uniquely identifying the dataset and checking data consistency according to the underlying ontological model. Finally, the data tier provides for persistent storage of user registration details, the storage of experiment meta-data in addition to unstructured and semi-structured files relating to the experiment, such as protocol documents and data files. Regarding the data files, the Open Data Initiative proposes the eXtensible Event Stream (XES) standard [3] as the preferred format for event data files, enabling a richer description of datasets through the attribution of semantics to data file (log file) content (Figure 3).

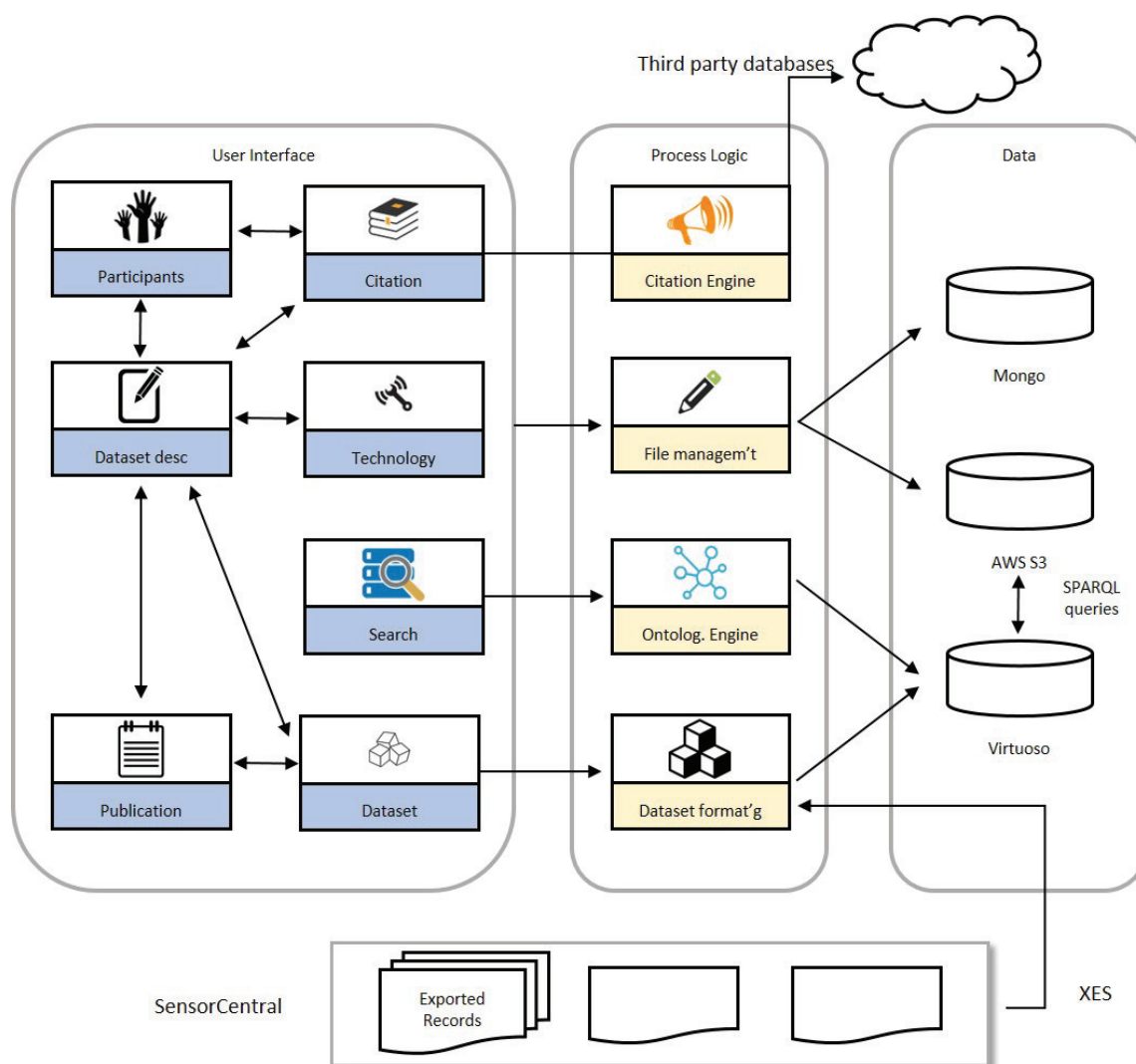


Figure 2. The ODI’s three-tier architecture—logical system view.

### 3.1. ODI portal—User Interface Components

In addition to datasets, the ODI will store experimental meta-data such as participant details, experiment protocol, ethical considerations and the types of activity being investigated. Informed by the ontology from previous work [2] for describing this information, the ODI portal user interface tier consists of the following components:

**Participants and Technology:** These components describes the sensor devices used to collect data from the participants involved in the study, enabling interpretation of data analysis outcomes.

**Publication and Citations:** These components capture existing citations of the datasets and any research papers describing the experiment, its data or results.

**Dataset:** This identifies the format of the uploaded dataset and facilitates the submission based on this, for example, CSV or XES formats.

**Search:** This component enables a user-friendly and flexible mechanism to build queries for retrieving existing datasets of interest to the researcher.

**Dataset description:** Data captured in this component is concerned with housekeeping such as user authentication, dataset registration, and a facility for ongoing retrieval and editing of previously uploaded datasets. Information from the user interface tier is posted as JSON query objects to the business logic tier of the platform.

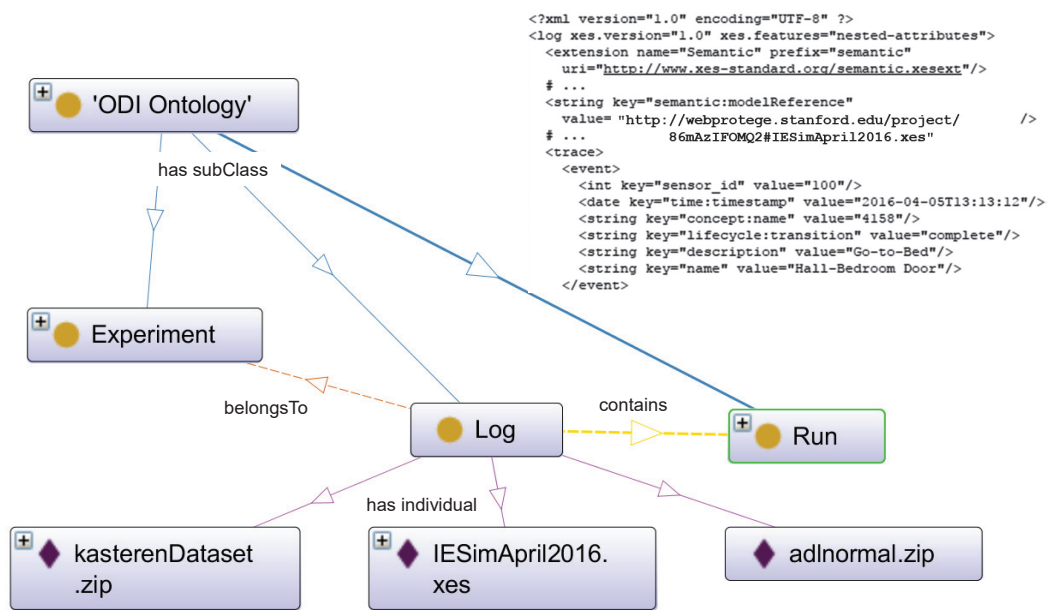


Figure 3. Adding semantics to data file (Log) content through XES.

### 3.2. ODI Portal—Process Logic and Data Tiers

The ODI portal server implements the data tier along with the processing features necessary to support secure and efficient data storage. Its main components are:

**File management:** This provides the storage of the experiment data files, along with any related documentation submitted as part of the upload such as protocol documents and ethical approval. This will be managed using a file store such as AWS S3.

**Citation engine:** This component provides citation data for the ODI datasets onto third party services. It uses public API resources such as those available on Scopus to monitor citations made towards ODI resources.

**Dataset formatting:** This component connects with the JSON query objects submitted from the user interface tier to populate, edit or query the data stored in the ontology database. It can also receive datasets exported in JSON or XES format from SensorCentral [9]. This will be achieved through SPARQL updates against a Virtuoso database.

**Ontological engine:** This component services the SPARQL queries submitted by the user.

### 3.3. Portal User Interface Prototype

The design philosophy for the front-end relies on adopting a user-friendly approach by enabling techniques to automatically fill out data fields, where possible. The aim is to provide an engaging experience so as to promote the use of the ODI (e.g., Figure 4). The prototype includes sections to group information by affinity in which the ontological structure that drives the ODI remains transparent to the end-user; bringing an intuitive experience, in which no system knowledge is required (e.g., Figure 5). Moreover, since not every pervasive health care study requires the same set of descriptive fields, the portal should dynamically include data fields depending on upload type. For example, whether the data file format is XES or CSV. Finally, as part of the citation promotion section, each dataset should be accessible through its own system generated Uniform Resource Identifier (URI).

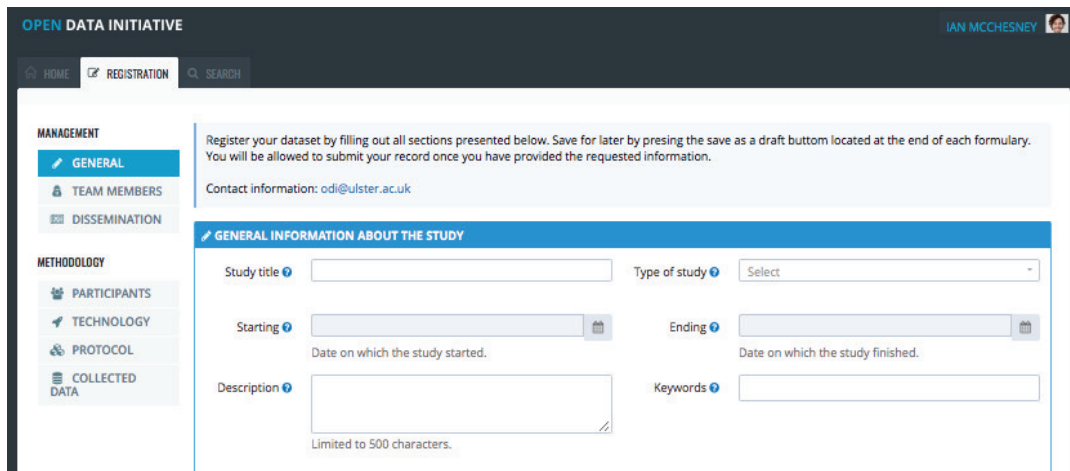


Figure 4. ODI portal sample screen for defining general study parameters.

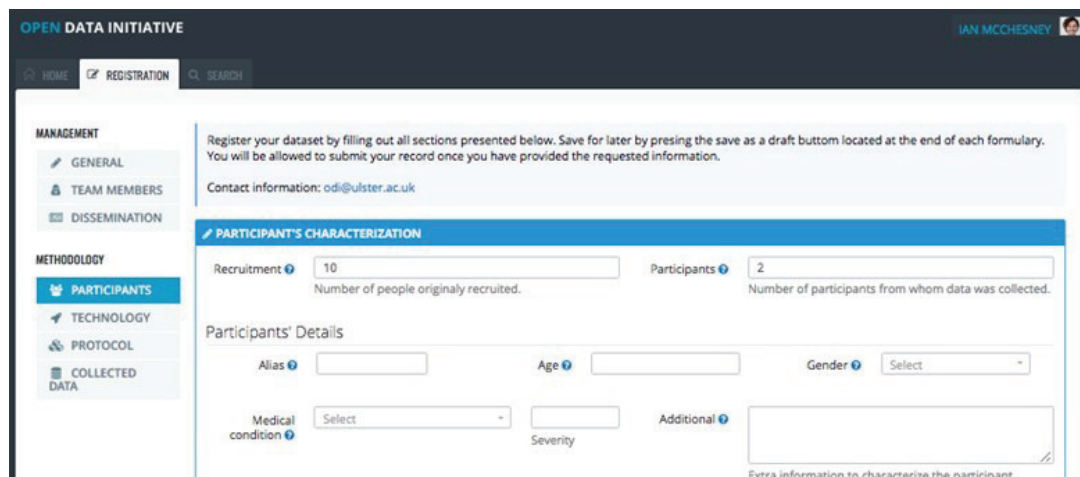


Figure 5. ODI portal sample screen for detailed capture of experiment participant information.

#### 4. Technical Review

The front-end prototype was made available to three experienced researchers in the field of activity recognition. They were given a simple protocol to follow for describing a recent experiment in which they were involved and preparing the upload of the associated datasets. On completion of the protocol tasks, a semi-structured group interview session was conducted, led by two of the authors and involving the three experienced researchers. The intervention was guided by the authors and technical notes were taken during the session. Views were sought regarding the prototype both in general terms and also in detail for each specific user interface section. The session lasted approximately two hours. While the number of evaluators is small, the purpose of this preliminary evaluation is to assess the adequacy of the current ODI data model before it is coded up as a functional portal system. It is envisaged that more formal evaluation of later iterations will involve a greater number of users to identify errors [10]. The technical review outcomes can be summarized under the broad headings of data entry considerations and dissemination considerations. The review also established that there was general support amongst the review participants for the concept and aim of the ODI project.

#### 4.1. Data Entry

When uploading a dataset, the user can designate their data file format as XES, CSV or other. When XES is used, the ODI portal will attempt to use the content of the event log to infer the types of activities being investigated in the study. The user would then be asked to confirm the activities before final upload. However, the absence of data fields which allowed the user to indicate generic activity types from the experiment (e.g., walking, making a drink, having a shower) was felt by the review participants to be an important omission in the experiment meta data. When querying the ODI, another researcher may be interested in seeking experiments relating to these generic activity types (e.g., find all datasets concerned with sitting AND walking), in addition to searching for specific activity descriptions as contained in the event logs.

From a usability perspective, greater flexibility was required during data entry. While the ODI aims to collect sufficient information to help a future researcher establish the quality and relevance of a dataset, the technical review participants all reported that the current information requirements were too rigid. Different datasets are collected under different organizational and experimental conditions. Prompting for too much meta-data (e.g., details of all team members, experiment participant data, device information) would not be meaningful in all situations. Overall, there should be more flexibility in population of the data fields, leading to overall less screen clutter.

The ability to upload other media to help describe the dataset would be useful. This could include, for example, a video recording of a participant completing an activity, a diagram showing the layout of a living space in which the experiment was conducted or a picture showing the specific placement of sensors.

The current user interface prototype allows the dataset owner to indicate the environmental conditions under which the data was collected (e.g., natural or controlled settings) and also to indicate the quality of the data. The technical review participants commented that even a poor-quality dataset can be useful to other researchers, which is consistent with findings reported in the literature [11]. For example, noise in data can be a useful feature when testing the limits of a future algorithm.

XES is an optional upload format. Since this is not yet widely used in the pervasive health community, the use of XES terminology in the user interface (e.g., the notion of “runs” or “traces”) should be removed where possible.

Depending on the type of experiment which is being uploaded, there should be the option to indicate if a dataset consists of labeled or unlabeled activities. To maximize uptake across the research community, there should be consideration for a multi-lingual user interface.

#### 4.2. Dissemination

There should be provision for the user to specify what are the dissemination requirements of the data. This includes guidelines for citing the work and any licensing requirements for future use. This may involve stipulation of Creative Commons Licenses such as CC-BY and CC0 [12], depending on the type of material intended for dissemination.

The inclusion of publications in which the dataset has been previously described is a valuable feature for evaluating the pedigree of the dataset and is useful for informing its potential value.

### 5. Proposed Revision and Concluding Comments

The ODI portal is informed by the data requirements as set out in the ODI ontology [2]. The technical review feedback thus has implications for the underlying data model as well as revisions to the portal's front-end.

The scope of revisions required can be illustrated through support for two broad usage scenarios. Consider John who wishes to use ODI for a flexible or lite upload of their data set. This will still require sufficient description to enable quality control before release. However, information such as device descriptions, participant profiles and research team membership, will be described in general terms (e.g., number of participants, device



types). By using ODI, John can quickly share his experimental work with the research community and obtain an ODI URI that he uses in his papers.

In the second usage scenario, consider Jane who is a multidisciplinary researcher from both medical and engineering fields. *She wishes to make available the dataset from a Randomized Control Trial (RCT) study in which five participants with severe Parkinson's conditions were involved. Once she received ethical approval and participant approval for sharing the collected data, Jane registers the sensor data collected providing detailed description of participant characteristics to distinguish between controlled and patient activities. Moreover, due the sensibility of the patients' conditions, she scrupulously provides information of the device characteristics (e.g., sensor type, data rate, software and hardware version) to facilitate respective data interpretation.* Thus, the ODI portal facilitates a more complete description of an experiment and its dataset for informed use and evaluation in other contexts.

Based on these outcomes, a revised version of the ODI ontology and portal is planned, followed by evaluation using a wider set of researchers from the pervasive health research community.

**Acknowledgments:** This research has been partially supported under the REMIND project Marie Skłodowska-Curie EU Framework for Research and Innovation Horizon 2020, under Grant Agreement No. 734355.

## References

1. Nugent, C.; Cleland, I.; Santanna, A.; Espinilla, M.; Synnott, J.; Banos, O.; Lundström, J.; Hallberg, J.; Calzada, A. An initiative for the creation of open datasets within pervasive healthcare. In Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, Cancun, Mexico, 16–19 May 2016; pp. 318–321.
2. McChesney, I.; Nugent, C.; Rafferty, J.; Synnott, J. Exploring an Open Data Initiative Ontology for Shareable Smart Environment Experimental Datasets. In *International Conference on Ubiquitous Computing and Ambient Intelligence*; Springer: Cham, Switzerland, 2017; pp. 400–412.
3. IEEE. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. *IEEE Std 1849–2016* **2016**, 1–50, doi:10.1109/IEEESTD.2016.7740858.
4. McNiff, J.; Whitehead, J. *All You Need to Know about Action Research*; SAGE: Newcastle upon Tyne, UK, 2011.
5. Easterbrook, S.; Singer, J.; Storey, M.A.; Damian, D. Selecting empirical methods for software engineering research. In *Guide to Advanced Empirical Software Engineering*; Springer: Berlin, Germany, 2008; pp. 285–311.
6. Hayes, G.R. The relationship of action research to human-computer interaction. *ACM Trans. Comput.-Hum. Interact.* **2011**, *18*, 15.
7. Zuber-Skerritt, O. Action learning and action research: paradigm, praxis and programs. In *Effect. Chang. Manag. Action Res. Action Learn*; Sankara, S., Dick, B., Passfield, R., Eds.; Southern Cross University Press: Lismore, Australia, 2001; pp. 1–20.
8. Bourque, P.; Fairley, R.E. *Guide to the Software Engineering Body of Knowledge (SWEBOK (R)): Version 3.0*; IEEE Computer Society Press: Washington, DC, USA, 2014.
9. Rafferty, J.; Synnott, J.; Ennis, A.; Nugent, C.; McChesney, I.; Cleland, I. SensorCentral: A Research Oriented, Device Agnostic, Sensor Data Platform. In *International Conference on Ubiquitous Computing and Ambient Intelligence*; Springer: Berlin, Germany, 2017; pp. 97–108.
10. Borsci, S.; Macredie, R.D.; Martin, J.L.; Young, T. How many testers are needed to assure the usability of medical devices? *Expert Rev. Med. Devices* **2014**, *11*, 513–525.
11. Hernández, N.; Castro, L.A.; Favela, J.; Michán, L.; Arnrich, B. Data Quality in Mobile Sensing Datasets for Pervasive Healthcare. In *Handbook of Large-Scale Distributed Computing in Smart Healthcare*; Springer: Berlin, Germany, 2017; pp. 217–238.
12. Commons, C. Creative Commons—About The Licenses. 2018. Available online: <https://creativecommons.org/about/> (accessed on 28 May 2018).

