Software solutions for form-based, mobile data collection - A comparative evaluation

Markus D. Steinberg[1], Sirko Schindler[2], Friederike Klan[3]

Abstract:

Many citizen science projects rely on their contributors going to the field and collecting data. Due to their wide availability and increasing capability, modern mobile devices have become an indispensable tool to ease the collection process. Projects can publish mobile apps, that allow contributors to easily collect data and submit their results. The requirements of individual projects oftentimes overlap to a large extent, which triggered the development of multiple generic frameworks. They allow new projects to quickly generate customized apps and reuse existing infrastructure. However, the wide landscape of tools with diverging capabilities requires projects to compare and choose. This report supports data managers in making an informed decision. We report on our experiences primarily on the whole data collection workflow starting from setting up your own instance to finally analyzing the retrieved data. We compare eight tools – both free and commercial – according to the features provided and difficulties encountered.

Keywords: Data collection; Citizen Science; Mobile Software; Web Applications

1 Introduction

Collecting field data is an integral part of many projects in citizen science and other fields of research, especially the life sciences. Mobile applications offering form-based-surveys and making use of built-in sensors are increasingly used to facilitate this process. Since the creation of mobile field surveys from scratch can be a tedious task, multiple tools have been developed that can help with form design, data collection via mobile devices, data export and storage. Some even support simple data analysis.

With the increasing number of software options, data managers often face the question, which tool is most suitable for their current use case. To provide a clear foundation for such a decision, this paper evaluates and compares a selection of software tools for survey design

[1] Friedrich Schiller University Jena, Institute for Computer Science, Ernst-Abbe-Platz 2-4, 07743 Jena, Germany markus.daniel.steinberg@uni-jena.de

[2] German Aerospace Center (DLR), Institute of Data Science, Mälzerstraße 3, 07745 Jena, Germany sirko.schindler@dlr.de, `https://orcid.org/0000-0002-0964-4457`

[3] German Aerospace Center (DLR), Institute of Data Science, Mälzerstraße 3, 07745 Jena, Germany friederike.klan@dlr.de, `https://orcid.org/0000-0002-1856-7334`

and mobile data collection. To the best of our knowledge no previous, comparable study on this topic has been published yet.

All tools were tested with respect to different aspects relevant to their usage. A list of such tool-characteristics was compiled from the features that are advertised by the tools themselves. Availability and usability of these features were then evaluated by thoroughly testing each of the tools: First, general information about the tool like its open source repository (if available) and its license were identified by consulting its website and documentation. Then, if no web-based version was offered, a local, self-hosted instance of the tool was set up. Afterwards the typical data collection workflow (Fig. 1) was executed: (1) a form was designed, examining all available form-elements and form-building features like skip-logic or localization, (2) the survey was deployed to the mobile app, (3) sample data were collected, (4) submitted to the server and then (5) exported. Finally, (6) additional features like visualization or data encryption were explored. In cases where the availability of features was not obvious, the tool's community channels or its support team were consulted for clarification.

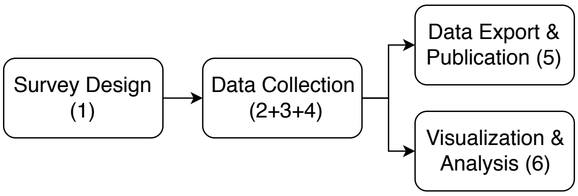

Fig. 1: Data collection workflow

Of course, such a study will never be able to cover all published tools, especially since their number is ever growing. The examined tools were selected because of their active and rapid development, their large user community, explicit recommendations by the tool's users, their extensive feature repertoire, or their professional design.

The following tools were considered:

- EpiCollect5 (EC5)[4], developed at the Imperial College London, the successor of Epicollect [1] and Epicollect+ (Plus)[5],
- Open Data Kit 1 (ODK1) [5] and Open Data Kit 2 (ODK2) [3], the two open source tool suites that are being developed by Nafundi[6] and the ODK community,
- Ohmage, an open source data collection platform that promises additional features like data analysis and visualization, developed at the University of California, Los Angeles[7], and Cornell Tech [8] [10, 13],

[4] https://five.epicollect.net
[5] http://plus.epicollect.net/
[6] https://nafundi.com
[7] http://www.ucla.edu/
[8] https://tech.cornell.edu

- KoBo Toolbox, an open source data collection tool being developed by members of the Harvard Humanitarian Initiative (HHI) [6],
- SurveyCTO [12] and Magpi [8], two paid subscription-based data collection platforms, which also offer free subscriptions that come with a few restrictions. For this study, the capabilities of the free versions were examined.

The Fieldtrip Open software suite developed by the EU-project Citizen Observatory Web (COBWeb), described in [7] and published in [4], was initially considered as well, but was deemed unfit for thorough testing (see details in Sect. 2). The data collection tool BioCollect, which is popular in the biodiversity domain and was developed by the Atlas of Living Australia [2], was not examined in this study because, in contrast to the other examined tools, it is not generic, but tailored to recording species observations.

Sect. 2 reports on our hands-on experiences with the examined tools in the different phases of the data collection workflow and compares their features. Sect. 3 highlights the most important results and makes suggestions for future development and improvement of mobile data collection tools.

2 Comparative evaluation

The comparison evaluates the conditions for use and customization set by the considered tools and frameworks, discusses installation related aspects and goes on with usage related features grouped together according to the individual steps of the data collection workflow as illustrated in Fig. 1. A more in-depth version of this evaluation was published in [11].

Conditions for use and customization One of the first aspects that should be considered in the decision for or against a certain tool is the conditions it sets for using and extending it. The following facets were examined. The results for these elements are shown in Tab. 1.

Active Development Is the software currently under active development, i.e. can we expect that new features will be added over time and software bugs are fixed? This is judged, if possible, by commits in the past six months, otherwise by activity on social media or in forums.

License Under which conditions can the software be installed, used, or extended?

Open Source Is the source code of the software provided as open source? This includes all parts that are required to build and deploy a survey, collect data with a mobile device and store the data on a server.

Programming language Which programming languages are used for developing the tool?

Self-hosting Is it possible to host the software yourself, so the collected data is stored on your own server?

Tab. 1: Usage and development related criteria.
(●. . . criterion fulfilled; ○. . . criterion not fulfilled)

	EC5	ODKv1	ODKv2	Kobo	Ohmage	SurveyCTO	Magpi	COBWEB
Active Development	●	●	●	●	○	●	●	○
Open Source	○	●	●	●	●	○	○	●
Programming language	-	Java JavaScript Python	Java JavaScript	Java JavaScript Python	Java Objective C	-	-	JavaScript Python
License	-	Apache[a]	Apache[a]	Apache[a], GNU[b]	Apache[a]	-	-	BSD3[c]
Self-hosting	○	●	●	●	●	○	○	●

[a] Apache License 2.0 [b] GNU Affero General Public License v3.0 [c] 3-Clause BSD License

Software installation and technical issues Some of the examined tools have components that need to be installed by the form author before they can be used. As an outcome of this step, we decided to exclude the Fieldtrip Open software suite from futher testing. Due to the missing documentation for its different parts, the software could not be fully set up in order to properly test its capabilities. Multiple attempts to set up the required persistence middleware on a Windows OS were unsuccessful. The middleware and the survey designer could finally be run on a Linux based computer, but the button that is supposed to save the designed survey resulted in JavaScript errors hinting at (1) access control problems and (2) problems inside the running middleware. It is unclear whether these errors occurred due to software problems or due to incorrect or missing configuration (which, in turn, could be the result of missing documentation).

The ODK suites require form authors to set up a server, ODK Aggregate, on a cloud-platform (AWS, Google, . . .) or host it themselves. This server is then used for survey distribution, data storage, visualization and data export. ODK2 additionally requires the setup of the ODK Application Designer which in turn requires Java, Google Chrome, NodeJS, Grunt and the Android SDK as prerequisites.

The other tools provide a web-based UI and do not involve an installation. Thus, no technical knowledge is required for their usage. All tools that offer a self-hostable version of their software provide detailed guides for the required steps. KoBo even maintains a Docker-version for a convenient setup. SurveyCTO's support team also mentioned in an email conversation that self-hosting of their software could be arranged on a case-by-case basis for their paying customers, provided that all additional costs will be settled by the user.

Survey design Almost all of the studied tools offer a form designer, a graphical user interface facilitating survey design, for example by allowing to add and arrange form elements via drag and drop. In the following, features are described that simplify form-authoring or allow to build more sophisticated and user-friendly surveys. The respective support among the tools is shown in Tab. 2.

Skip Logic Skip certain parts of a survey depending on previous answers.

Localization Define labels for questions in multiple languages, so the survey can automatically be translated to a user's preferred language.

Calculations Evaluate mathematical or logical expressions referencing answers to preceding questions in a survey and use the results in skip logic, text-blocks, etc.

Queries Read data from a structured source (e.g. a CSV file) and use the results for skip logic, as answer-options, etc.

Linked Tables Launch subforms that store data in different database tables.

Required & optional fields Mark a question as mandatory or optional to indicate whether the survey can be finished without providing an answer.

Validation Define validity constrains for form-fields, e.g., the range of valid values for a number input.

Building custom prompts Build prompts with custom functionality, typically using a markup language like HTML for the presentation and a programming language to define the functionality.

Tab. 2: The survey design features provided by the examined tools.
(●. . . feature included; ◐. . . feature partially included; ○. . . feature not included)

	EC5	ODKv1	ODKv2	Kobo	Ohmage	SurveyCTO	Magpi
Form Designer	●	●	○	●	●	●	●
Skip Logic	●	●	●	●	○	●	●
Localization	○	●	●	◐[a]	○	●	○
Calculations	○	●	●	○	○	●	●
Queries	○	○	●	○	○	○	○
Linked Tables	○	○	●	○	○	○	○
Required & Optional Fields	●	●	●	●	●	●	●
Validation	●	●	●	●	●	●	●
Building custom prompts	○	○	●	○	○	○	○

[a] Not supported in form designer, has to be added manually by exporting the form, editing the .xls file and then importing it.

The examined tools support and guide inexperienced form authors to a different extent. Particularly noteworthy are the extensive help-texts that are provided by SurveyCTO. They explain the typical workflow in the user interface as well as the different features and options that are available for each step. Inexperienced authors are thus guided through all necessary steps. SurveyCTO and Magpi also offer template forms that can be used to learn about the different form elements and their configuration. ODK1 provides explanations for the configuration of available form elements but the workflow-guidance is missing in the user interface. The reason is most likely the fact that ODK's form builder is a tool that is separated from the deployment-server and therefore also has a separate user interface.

Worth mentioning are also the wizards that EpiCollect5, KoBo, SurveyCTO and Magpi offer to build, for example, skip logic expressions. These wizards greatly simplify the formulation of complex logic statements, especially for inexperienced form authors.

As shown in Tab. 2, ODK2 offers some additional features like queries, linked tables and custom prompts that none of the other tools provide. However, ODK2 is the only one of the examined tools that does not offer any kind of form designer. Forms have to be created as .xls files and are then transformed using the ODK Application Designer. This, in addition to the more complex and technical deployment of surveys to the server and then to mobile devices, makes the usage of ODK2 more difficult compared to other tools. Thus, as the official ODK help page emphasizes, the usage of ODK2 is only recommended if the additional features are required for a certain use case [9].

The form design also depends heavily on the input elements that are available. All of the tools provide support for the most basic types of information: text input, integer and decimal numbers, dates and times as well as single- and multiple-choice questions. They also allow to display textual information to the data collector.

Location information, which is a very important factor in many surveys, is also supported by all examined tools in the form of automated location determination using the collection device's sensors. However, manual input of a location, for example by placing a pointer on a map or explicitly stating latitude, longitude, altitude and/or accuracy, is only supported by ODK1 and SurveyCTO. More complex geographical information like paths (a sequence of locations) or an area (a closed path) are only supported by KoBo and SurveyCTO.

Images can be collected by all tools except Magpi; audio recordings, video recordings and barcodes by all except Magpi and Ohmage. Additionally, KoBo and Magpi offer some unique input elements: KoBo is currently developing and testing ratings (e.g. assigning "good", "bad" or "neutral" to a defined set of options) and rankings (ranking a predefined set of options). The latter could, for example, be used to express personal preferences, e.g. allow the user to state that he prefers apples over bananas over oranges.Magpi on the other hand allows to read information via Near-Field Communication[9].

[9] http://nearfieldcommunication.org/about-nfc.html

Data collection After the survey is designed and deployed on a server, data can be collected via a mobile device, typically using a mobile app. All of the examined tools provide support for offline data collection, meaning that data can be collected without an active internet connection and can later be submitted.

The support for mobile operating systems differs among the tools: All of them support Android-OS but only EpiCollect5, Ohmage and Magpi also offer apps for iOS. Apart from these two, no other operating systems are supported. For the open source tools, we were able to verify that the data collection apps are developed as native apps and not as cross-platform mobile applications. However, for closed source projects we were not able to obtain this information.

Another important factor in the data collection step is, if metadata are automatically gathered and stored with the collected data. Metadata describe the context of the collection process and the individual data records and therefore allow for a meaningful interpretation of the data. All tools provide information about date and time of data collection or submission as well as some kind of identification of the user who collected the data (username or e-mail address). ODK1, KoBo and SurveyCTO also allow to collect information about the mobile device that was used: the device-ID (IMEI), the subscriber-ID (IMSI), the SIM serial number or the phone number.

In cases where highly sensitive or private data is collected in a survey, the form author might want to ensure that no one will be able to get unauthorized access to the data. For such cases ODK1, KoBo and SurveyCTO provide the option to store a public key that automatically encrypts the data as soon as it is saved on the mobile device. The data can only be decrypted with the matching private key once the data is downloaded from the storage server. This ensures that the data is not only secure during the submission (which would usually be assured via the SSL/TLS protocol) but also while it is stored on the server, thus providing a strong level of security.

With automated quality checks, SurveyCTO offers an feature that can drastically improve data quality. In addition to the validation of the values inserted into form fields, such checks allow to detect submitted values that are outliers, values that occur too frequently, or other potentially faulty items. These quality checks can be configured to be run at certain time intervals and to be reported to a given e-mail address.

Data export & publication Regardless of the data collection tool, once the collected data are uploaded to the server, they can be exported as a file. The available file types differ among the examined tools: All of them support data export in the form of CSV files. As seen in Tab. 3 most of them also support some other file types that can be useful depending on the use case or user preferences.

Beside data export, data publication is directly integrated in some of the tools. EpiCollect5 provides an API to access its data and a guide on the usage of that API to publish the collected data to Google Spreadsheets. ODK1 offers direct data publication to Google Spreadsheets, Google FusionTables, REDCap[10] servers and custom JSON servers. SurveyCTO also supports export to Google Spreadsheets and Google FusionTables. It additionally offers an integration with Zapier[11]. This is especially noteworthy since Zapier is a platform that can be used as a "bridge" to integrate the published data with hundreds of different services and applications.

A feature currently not provided by any of the examined tools is the semantic enrichment of the collected data. Embedding the assembled data in a semantic framework and interlinking individual data items with one another can give further interpretational context and allows a more seamless integration with other projects or information sources.

Tab. 3: The data export formats supported by the examined tools.
(●. . . feature included; ○. . . feature not included)

	EC5	ODKv1	ODKv2	Kobo	Ohmage	SurveyCTO	Magpi
CSV	●	●	●	●	●	●	●
JSON	●	●	○	○	○	○	○
XLS	○	○	○	●	○	●	●
XML	○	○	○	○	○	○	○
XML/KML	○	●	○	○	○	●	○
RDF	○	○	○	○	○	○	○

Data visualization & analysis All of the tools provide some kind of built-in support for data visualization, though the supported types vary as seen in Tab. 4. On the other hand, data analysis is currently not supported by any of the tools. Ohmage claims to provide such features but does not properly integrate them in the tool's interface. SurveyCTO provides a way to monitor the incoming data and visualize relationships between different fields of data, but does not offer the capabilities of full analytic software. Magpi seems to have a similar feature as SurveyCTO but it is locked for free users.

Currently, the best way to analyze the collected data, regardless of the tool, is to either export it using one of the supported file types and then import that file in a data analysis tool of one's choice or to publish the data on a cloud-based platform and then use analysis tools that integrate well with the platform.

[10] https://www.project-redcap.org

[11] https://zapier.com

Tab. 4: Visualization types, analysis and semantic enrichment features
(○. . . feature not included)

	EC5	ODKv1	ODKv2	Kobo	Ohmage	SurveyCTO	Magpi
Visualization	Map, Pie chart	Map, Pie chart, Bar chart	Map	Map, Pie chart, Bar chart, Line chart, Area chart	Map, Pie chart, Bar chart, Line chart	Map, Pie chart, Bar chart, Scatterplot, Trend plot[a]	Map
Analysis	○	○	○	○	○	○	○
Semantic Enrichment	○	○	○	○	○	○	○

[a] Plotting a numeric value over time

3 Conclusion

The presented comparison shows that due to the different features that are offered by the different software tools, the choice of a platform depends on the given use case with its unique requirements. However, it also shows that ODK1 and KoBo Toolbox are the open source tools that offer the most comprehensive set of features. SurveyCTO, on the other hand, offers the most professional and user-friendly environment if the limitations of the free subscription are of no concern. For most data collection projects, at least one of these three tools should be able to cover the requirements.

Another point that this comparison shows very clearly is that none of the tools currently provide any kind of semantic component that would provide a unique and machine-comprehensible semantics of the exported data. The only tool that seemed to take a step in this direction was the COBWeb software suite with its RDF export, which was not fit for proper testing. Since linked and semantically enriched data enable more meaningful interpretation of the data and even automated reasoning, such features could drastically improve both quality and usability of the collected data. Therefore, such features deserve some attention in the future enhancement of data collection platforms and tools.

Analysis support for the collected data is another point that could be integrated in the tools. Currently, projects have to rely on external tools for such features, which means that the data either has to be transferred to a cloud platform or manually exported and imported into some analysis software. Both options require additional time-consuming effort and could in some cases even create privacy issues if highly sensitive data is involved.

A third improvement that could be taken into account by the examined tools is the use of cross-platform technologies for mobile applications. Advantages here would be two-fold: Data collection projects would only have to maintain a single code base for all their mobile applications. At the same time this could increase the number of potential survey participants and therefore the engagement in projects that involve data collection.

References

[1] David M Aanensen et al. “EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection”. In: *PLoS ONE* 4.9 (2009-09). Ed. by Simon I. Hay, e6968. DOI: 10.1371/journal.pone.0006968.

[2] Peter Brenton. “BioCollect - A modern cloud application for standards-base field data recording”. In: *Biodiversity Information Science and Standards* 2 (2018-05), e25439. DOI: 10.3897/biss.2.25439.

[3] Waylon Brunette et al. “Open data kit 2.0: expanding and refining information services for developing regions”. In: *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications - HotMobile ’13*. ACM. ACM Press, 2013, p. 10. DOI: 10.1145/2444776.2444790.

[4] Citizen Observatory Web. *COBWEB - Software Outputs*. URL: https://cobwebproject.eu/news/publications/software-outputs (visited on 2018-10-31).

[5] Carl Hartung et al. “Open data kit: tools to build information services for developing regions”. In: *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development - ICTD ’10*. ACM. ACM Press, 2010, p. 18. DOI: 10.1145/2369220.2369236.

[6] Harvard Humanitarian Initiative. *KoBoToolbox: Data Collection Tools for Challenging Environments*. URL: https://www.kobotoolbox.org/ (visited on 2018-10-27).

[7] Christopher I Higgins et al. “Citizen OBservatory WEB (COBWEB): A generic infrastructure platform to facilitate the collection of citizen science data for environmental monitoring”. In: *International Journal of Spatial Data Infrastructures Research (IJSDIR)* 11 (2016), pp. 20–48. DOI: 10.2902/1725-0463.2016.11.art3.

[8] Magpi. *Advanced Mobile Data Collection, Messaging, and Visualization*. URL: https://home.magpi.com/ (visited on 2018-10-27).

[9] Open Data Kit. *ODK Help page*. URL: https://opendatakit.org/help/) (visited on 2018-10-28).

[10] Nithya Ramanathan et al. “ohmage: An open Mobile System for Activity and Experience Sampling”. In: *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE. IEEE, 2012, pp. 203–204. DOI: 10.4108/icst.pervasivehealth.2012.248705.

[11] Markus D. Steinberg. *Software solutions for form-based collection of data and the semantic enrichment of form data*. arXiv: 1901.11053 [cs.CY].

[12] SurveyCTO. *Collect data you can trust*. URL: https://www.surveycto.com/ (visited on 2018-10-27).

[13] Hongsuda Tangmunarunkit et al. “Ohmage: A general and extensible end-to-end participatory sensing platform”. In: *ACM Transactions on Intelligent Systems and Technology* 6.3 (2015-04), pp. 1–21. DOI: 10.1145/2717318.