Final Report: Interactive distributed corpus exploration and annotation infrastructure for large corpora and knowledge-bases (INCEpTION)

1 General Information

DFG reference number EC 503/1-1 and GU 798/21-1

Applicants

- Dr. Richard Eckart de Castilho, Department of Computer Science, Technische Universität Darmstadt
- Prof. Dr. Iryna Gurevych, Department of Computer Science, Technische Universität Darmstadt

Title of the project

- (de) Infrastruktur für interaktive verteilte Exploration und Annotation grosser Korpora und Wissensbasen
- (en) Interactive distributed corpus exploration and annotation infrastructure for large corpora and knowledge-bases

Internet addresses of the project

- https://inception-project.github.io
- https://github.com/inception-project
- https://www.informatik.tu-darmstadt.de/ukp/research_ukp/ukp_research_projects/ukp_project_inception/index.en.jsp

Period covered by the report 09. Nov 2016 - 30. Sep 2022

Overall funding period

- initially approved period 09. Nov 2016 - 30. Sep 2022
- cost-neutral extension granted 12.03.2018 (to 30. Sep 2021)
- cost-neutral extension granted 22.04.2021 (to 30. Sep 2022)

2 Final Progress Report

Background and objectives of the project The goal of INCEpTION was to create a modular and generic research infrastructure for corpus annotation that scales to large text document collections by flexibly building subcorpora. Originally, the target user groups were envisioned to be mostly computational linugists and corpus linguists. However, during the course of the project, a user community covering a much broader range of disciplines emerged, including but not limited to various Digital Humanities disciplines, empirical social sciences, and life sciences.
INCEpTION aimed to address the users' need to perform selective semantic annotation tasks within and across documents, and thereby to enable a phenomenon-based access to the huge amounts of digitally-available text. INCEpTION was to support expert users in exploring large document collections, in setting up annotation schemes, and in extracting task-specific subcorpora from a large background corpus. The annotation of the corpora was to be flexibly distributable to remotely-working annotation teams of different qualification levels and backgrounds. The work of these teams was to be supported through prioritisation and annotation suggestions based on machine learning technology (human-in-the-loop) to efficiently create a large corpus with high-quality annotations for training and evaluating the respective algorithms. Further functionality was to be provided to maintain and expand the knowledge bases used during the semantic annotation tasks as well as to connect to external standard knowledge bases.
The infrastructure was to be realized as a modular and extensible architecture and to be distributed as a liberally licensed open source software package.

Work steps during the reporting period, including deviations from the original plan; any organisational or technical problems Our strategy to make INCEpTION a successful and accepted annotation platform focused on a close interaction with the (potential) user community and the early integration of user feedback. During the first year of the project, we implemented the base functionalities in three areas:

- **Subcorporation:** The ability to search (large) external document repositories for documents relevant for annotation and the ability to search the annotated documents themselves using text and/or annotation patterns. We tried out different technologies and eventually settled on using MTAS[1] for the search on annotated documents. For access to external document repositories, we implemented a plug-in mechanism through which support for different types of repositories can be added. At the end of the funding period, four types of external repositories are supported: ElasticSearch, OpenSearch, Apache Solr and the PubAnnotation[2] service.
- **Annotation recommendation:** To improve the efficiency of the annotation process, we implemented the ability to leverage *recommenders*. These are internally-trained machine learning models or external machine learning services that generate annotation suggestions which users can accept, reject or correct. These models can be statically (pre-)trained or they can be dynamically re-trained in a human-in-the-loop workflow. Users can optionally employ an active-learning guided process to improve the learning efficiency of a human-in-the-loop workflow. Again, a plug-in mechanism can be used to support different types of machine learning libraries. At the end of the funding period, there are eight internal recommender implementations. There is a generic external recommender API that allows, e.g., to connect popular Python-based machine learning implementations including popular deep learning libraries. Finally, there is support for using external text analysis services such as those offered by the European Language Grid, CLARIN WebLicht, LAPPS Grid (deprecated) or Huggingface (experimental).
- **Knowledge management:** To enable the disambiguation of entities in texts, INCEpTION supports internal and external knowledge bases. We decided to build on semantic web standard technology: RDF/OWL and SPARQL. INCEpTION supports internal knowledge bases as well as the ability to connect to SPARQL services which are, e.g., freely offered by the ZBW Leibnitz Information Centre for Economics, DBPedia, Wikidata, the Food and Agriculture Organization (FAO) of the United Nations and many more. There are several server products on the market that implement the SPARQL protocol and each of them has their own quirks and specialties. In particular, we found that the ability to perform a fast full-text search (FTS) on the knowledge bases was essential for their interactive use in INCEpTION, but FTS is not standardized in the SPARQL protocol. It was

[1] https://github.com/textexploration/mtas
[2] https://www.pubannotation.org

very challenging and time-consuming to implement a smooth abstraction layer, but still we managed to support most major SPARQL service implementations at the end of the reporting period.[3]

The first public release of INCEpTION in April 2018, about 18 months into the project, included early implementations for all of the above-mentioned areas of functionality. Shortly before that first public release, we already organized a workshop that provided the participants with an early-access glimpse of the functionality and which allowed us to gather information allowing to prioritize further development.

As can be seen in the remainder of the project report, staying in touch with the user base was (and is) very important to us. Once a solid connection to the user base had been established, we departed from several ideas that were included in the original proposal.

A major difference between the proposal and the actually implemented INCEpTION platform is related to using a cluster system to automatically analyze large text corpora. With the fast rise of deep learning early in the project, the focus of the community shifted away extremely quickly from high-performance cluster computing towards GPU-based deep learning. The ability to connect state-of-the-art algorithms as annotation recommenders to INCEpTION was more important than the ability to apply less-sophisticated algorithms to large data using compute clusters. Thus, we dropped the idea of supporting cluster computing and instead focused on the ability to interact with GPU-based computing via a Python-interoperability layer and cloud services like the European Language Grid, LAPPS Grid, CLARIN WebLicht or Huggingface. Likewise, the idea to implement sophisticated user interfaces for the configuration of individual machine learning algorithms was dropped. The integrated machine learning algorithms come with simple configuration options to make them easily accessible by non-expert users. Users who are experts in machine learning can alternatively connect their own models through the external recommender API, giving them maximum choice and configuration flexibility for their machine learning approaches.

Also, we had originally assumed that the ability to perform focused annotation would be more important to our users than it ended up being. By focused annotation, we mean the ability to search in a large corpus in order to selectively annotate it using a special annotation view that does not display the entire document, but only the context of each search result. The ability to search through external document repositories was accepted well by users, and for some was a key reason why they chose to use INCEpTION over other tools. The ability to search over annotated documents within INCEpTION was also important for many users. However, the extra step of creating a focused annotation editor showing only snippets of documents for annotation turned out to be unnecessary. Instead, the internal and external search functionalities were refined and better integrated with the kinds of annotation workflows our users needed.

Experience regarding methods employed and reuse options With respect to the methods employed for building the INCEpTION annotation platform, we started off using the modules from the code of the earlier WebAnno open source annotation editor, extended the modularity of its architecture and added new INCEpTION-specific functionalities: internal and external search, knowledge-base support and a flexible annotation recommendation subsystem (just to name a few). The modular architecture often allowed us to implement and evolve the different functionalities of the platform at different speeds, depending on what was important for our actual users at the time. However, a significant amount of time had to also be invested into refactoring, to ensure that the architecture remained modular despite its continuous growth as new levels of abstractions had to be introduced. The INCEpTION software was built on the WebAnno code base, which consisted of 39 modules when our project started. This number more-than-doubled to **100+ modules** in the INCEpTION code base.[4]

What the modular architecture does for INCEpTION internally, the external APIs and support, for standard protocols like SPARQL do in terms of interoperabilty with other systems. The INCEpTION remote API is based on the OpenMinTeD AERO protocol[5]. It supports setting up projects, monitoring projects and exporting project data, as well as being notified about changes in the project state at the level of

[3] Apache Jena Fuseki, Eclipse RDF4J, Stardog, Virtuoso and the Wikidata entity search service as supported by INCEpTION.

[4] At the start of the project, we had directly used many of the WebAnno modules and contributed many improvements made for the purpose of building the INCEpTION software back to the WebAnno code base for a long time. However, at some it became clear that we needed to perform several fundamental refactoring steps to further develop the INCEpTION architecture and backporting these steps to WebAnno would have been very time intensive. At this point, it was decided to absorb those parts of the WebAnno code base that we used directly into the INCEpTION code base and to perform the refactorings only within INCEpTION. Due to this joint history of WebAnno and INCEpTION, most users of WebAnno have an easy upgrade path to INCEpTION and can import their WebAnno projects directly into the new platform.

[5] https://openminted.github.io/releases/aero-spec/1.0.0/omtd-aero/

individual annotators, documents or the entire project. This enables the annotation functionality offered by INCEpTION as a component within a larger workflow.
Please refer in particular to the *Featured use-cases* in the *Response* section further ahead in this document for an illustration of the wide range of tasks and domains that the module INCEpTION text annotation platform can be used for.

Results

The result of the project is the INCEpTION software. This is an open source software package licensed under the Apache License 2.0 which users can install on their own personal hardware, on servers, or in the cloud. While we do host some instances of INCEpTION, these instances are explicitly not offered as infrastructure services to the research community. Local, regional or national research infrastructure providers (e.g., CLARIN-EL) may offer INCEpTION instances to their respective users and may also promise a certain level of service availability. INCEpTION is not such a research infrastructure provider. We provide the software to the research infrastructure providers.

Releases and downloads The INCEpTION software is available for download from the INCEpTION homepage[6] and from GitHub[7] as of April 2018. The first public release was **v0.2.2** (April 2018); the current release is **v24.1**. The first non-alpha release was v0.7.0 (Jan 2019) and there were a total of **79 stable public releases** since then (as of 30 Aug 2022, excluding beta releases, release candidates and two revoked released). Since the start of the project, we have had close to **14000 downloads** of release artifacts from GitHub. DockerHub[8] reports **100k+ pulls** of INCEpTION (as of 30 Aug 2022).

Documentation INCEpTION comes with a comprehensive built-in user manual which covers all aspects of regular use of the software. This user manual, as well as an administrator guide, are also available on the INCEpTION homepage. The administrator guide covers topics such as the installation of INCEpTION on a server, external authentication, upgrades, integration with third-party systems through the remote API, etc. There is also a basic documentation for developers that covers how to set up a development environment for INCEpTION, as well as certain parts of the architecture and the protocols for communication with external services.

Demo instance There is also a demo instance[9] of INCEpTION which can by used to try out the tool. It is a sandbox that can be accessed by anybody with publicly-known credentials. This instance is not meant for productive use. It may be shut down and its data may be reset at any time at our discretion.

UKP Lab instances We host an instance of INCEpTION at the UKP Lab. This is mostly used internally by the research group, but we have also occasionally created accounts for external collaboration partners and for use during workshops and tutorials. We also host a second *community* instance of INCEpTION, which can in principle be used by anybody with a CLARIN-AAI-compatible account (Shibboleth/SAML). This was set up mainly for the purpose of performing academically-crowdsourced annotation studies, but has also occasionally been used to provide external collaborators with a hosted instance. Users use these instances at their own risk and are asked to make regular backups of their data. UKP Lab does not offer any service-level guarantees for these instances.

Community contributions

The goal of every open source project is to attract community contributions. It is also one of the most difficult aspects of open source development. According to GitHub's *Insights* statistics[10], INCEpTION has had contributions from 46 different sources (until 23 Aug 2022). Most of these (28) were trivial/one-time contributions with 10 or less commits fixing small bugs, contributing documentation, etc. Most substantial contributions naturally originate from our project staff. Particularly notable contributions came from:

- (early 2019) the VISTA group at the Information Sciences Institute at the University of Southern

[6] https://inception-project.github.io
[7] https://github.com/inception-project/inception/releases
[8] https://hub.docker.com/r/inceptionproject/inception
[9] https://morbo.ukp.informatik.tu-darmstadt.de
[10] https://github.com/inception-project/inception/graphs/contributors?from=2016-11-09&to=2022-08-23&type=c

California[11] has been using INCEpTION to enable rapid adaptation of event detection and event argument attachment algorithms to new languages and ontologies.[12] To improve the user experience for this task, they contributed code for a new sidebar for the annotation page which allows the annotator to search an external document repository directly from the annotation page and to directly import and open such documents through the sidebar.[13,14]

- (late 2019) a course on software evolution and maintenance which was part of a master's program in software engineering[15] at Blekinge Institute of Technology held by Deepika Bapampudi and Michael Unterkalmsteiner. According to them, 62 students participated in the course. They told us that INCEpTION as an open-source project was perfect for inclusion in student projects, as they got open access to real-world projects. In addition, Michael Unterkalmsteiner had also previously contributed to INCEpTION which encouraged the inclusion in the course. The students evaluated the quality of the code using static analyzers and refactored the code smells, bugs, and vulnerabilities. The students first raised an issue related to a code smell, bug or vulnerability in the issue tracker, which was reviewed by the teachers, and upon approval, the issues were refactored and submitted through pull requests to the INCEpTION team. In addition, the students developed new features where they reflected on the analyzability of the INCEpTION code. We accepted some of these contributions.
- (mid 2020) the ENP China ERC project[16] not only used the external document repository search feature of INCEpTION, but they also contributed code[17] to support Apache Solr as an external document repository in INCEpTION.

Public relations

Hosted workshops and tutorials Originally, we planned to host two workshops. However, due to the Corona crisis that started around the time we would have hosted our second workshop, only the first workshop took place. For further talks, workshops and tutorials organized by third parties where we presented INCEpTION, please refer to the next section on response to public relation efforts.

- (12.-13. Mar 2018; half-day each) **FiF-Workshop 2018 INCEpTION**[18] was our kick-off, in-person workshop for the INCEpTION project in Darmstadt with external participants, even some from abroad. The workshop was used to introduce the ideas of INCEpTION and featured 7 short use-case presentations from the participants. In various discussion rounds, we discussed the ideas and use-cases and gathered further requirements and ideas for the project.
- (27 Nov 2018) **Linking Text and Knowledge using the INCEpTION annotation platform** (tutorial) for the Research Training Group GRK 1994: Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at TU Darmstadt. Presenters: Richard Eckart de Castilho, Naveen Kumar, Jan-Christoph Klie.

Posters without abstracts

- (21.-22. Sep 2017) **INCEpTION - A community-oriented smart semantic annotation platform**[19] (poster) at a workshop organized by the Digital Humanities Initiative in the RMU Network (DH-RMU) in Mainz. Presenter: Beto Boullosa.
- (05 Oct 2018) **INCEpTION - A Community-Oriented Smart Semantic Annotation Platform** (poster) at the Amazon Research Days, Berlin. Presenter: Richard Eckart de Castilho
- (30. Sep-02 Oct 2019) **Beyond WebAnno: The INCEpTION Text Annotation Platform**[20,21] (poster) at the CLARIN Bazaar of the CLARIN Annual Conference 2019, Leipzig. Presenter: Richard Eckart de Castilho.

[11] https://www.isi.edu/centers/vista/home

[12] https://inception-project.github.io/use-cases/rapid-events/

[13] https://github.com/inception-project/inception/issues/825

[14] https://github.com/inception-project/inception/pull/913

[15] https://www.bth.se/eng/education/masters/paaps20h/

[16] https://www.enpchina.eu

[17] https://github.com/inception-project/inception/pull/1720

[18] https://www.fif.tu-darmstadt.de/themen_fif/themen_details_32448.de.jsp

[19] http://www.rhein-main-universitaeten.de/news/get-together-der-initiative-digital-humanities-im-rmu-verbund

[20] https://www.clarin.eu/event/2019/clarin-annual-conference-2019-leipzig-germany

[21] https://www.clarin.eu/sites/default/files/clarin2019_bazaar_eckart_de_castilho.pdf

Posters with abstracts

- J.-C. Klie. INCEpTION: Interactive Machine-assisted Annotation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, pages 105–105, July 2018
- R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, and I. Gurevych. Linking Text and Knowledge using the INCEpTION annotation platform. In *Proceedings of the 14th eScience IEEE International Conference*, pages 327–328, Oct. 2018
- R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, and I. Gurevych. INCEpTION - Corpus-based Data Science from Scratch. In *Digital Infrastructures for Research (DI4R) 2018*, Oct. 2018

Peer-reviewed research or demo papers [22]

- (1 citation) B. Boullosa, R. E. de Castilho, A. Geyken, L. Lemnitzer, and I. Gurevych. A tool for extracting sense-disambiguated example sentences through user feedback. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, number TUD-CS-2017-0039, pages 69–72. Association for Computational Linguistics, April 2017
- (170 citations) J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018
- (4 citations) B. Boullosa, R. E. de Castilho, N. Kumar, J.-C. Klie, and I. Gurevych. Integrating Knowledge-Supported Search into the INCEpTION Annotation Platform. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing - Demo Papers*, pages 127–132, Aug. 2018
- (5 citations) R. E. de Castilho, N. Ide, J.-D. Kim, J.-C. Klie, and K. Suderman. Towards cross-platform interoperability for machine-assisted annotation. *Genomics & Informatics*, 17(2):e19., Juni 2019
- (1 citation) R. E. de Castilho, N. Ide, J.-D. Kim, J.-C. Klie, and K. Suderman. A multi-platform annotation ecosystem for domain adaptation. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 189–194, Florence, Italy, August 2019. Association for Computational Linguistics. Veranstaltungstitel: 13th Linguistic Annotation Workshop
- (26 citations) J.-C. Klie, R. E. de Castilho, and I. Gurevych. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. pages 6982–6993, Juli 2020. Veranstaltungstitel: The 58th annual meeting of the Association for Computational Linguistics (ACL 2020)
- (9 citations) M. Wu, N. S. Moosavi, A. Rücklé, and I. Gurevych. Improving QA generalization by concurrent modeling of multiple biases. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 839–853. Association for Computational Linguistics, Oktober 2020
- (2 citations) C. Scheunemann, J. Naumann, M. Eichler, K. Stowe, and I. Gurevych. Data collection and annotation pipeline for social good projects. November 2020. Veranstaltungstitel: AI for Social Good - AAAI Fall Symposium 2020
- (no citations yet) J.-U. Lee, J.-C. Klie, and I. Gurevych. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373, Juni 2022

Social media and mailing lists The projects maintains a public mailing list[23] which anybody can subscribe to and which users regularly go to to ask questions or post suggestions. As of Aug 2022, the mailing list has **208 subscribers**, and **over 1300 messages** in over **400 conversations** have been posted to the list. Releases are announced through the mailing list, as well as through GitHub's release announcement mechanism. Occasionally, major new features and new releases were also announced on Twitter.[24] The project maintains a YouTube channel where introductory videos have been published.[25]

[22] Citation counts were obtained from Google Scholar on 30 Aug 2022.

[23] https://groups.google.com/g/inception-users

[24] https://twitter.com/search?q=%23tap_inception

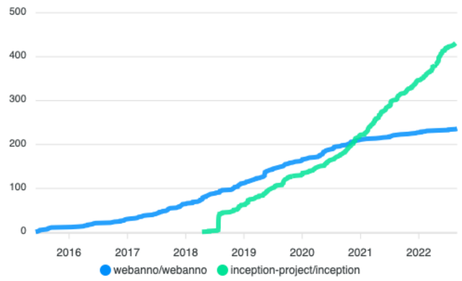
[25] https://www.youtube.com/channel/UC3sUTFFPYg0aWmZRag45yJw

User feedback The project maintains several public code repositories on GitHub.[26,27] These offer issue trackers where the project staff publicly manages their tasks. They also allow users to report issues and feature requests, or to ask questions (although questions usually reach us through the mailing list). The primary issue tracker of the project is in the main INCEpTION repository.[28] We have received a lot of feedback from users through this channel. Until Aug 2022, **over 160 users** (not counting members of the INCEpTION team) have opened a total of **over 400 issues** (bug reports, feature requests and questions). This feedback has helped us enormously in aligning the development of INCEpTION with the needs of the community. Because this feedback channel in particular has provided so much input already, we did not procure some of the additional channels we considered in the project proposal (WP 5), such as dedicated surveys or user studies. Including those reported by the team, at the time of writing there are over 1900 tracked issues, ≈1700 of which have already been resolved. These counts only cover the main repository, not the other repositories under the INCEpTION GitHub organization.

Additionally, anybody can contribute to the project by opening a pull request (PR). PRs are reviewed by the project maintainers and are merged if useful, possibly with revisions. For more information, please refer to the section *Community contributions* above.

Has there been any response to the publicity for the project?

Over the course of the project, we have observed steadily increasing interest in the INCEpTION platform. We have had many contacts with all kinds of users: often individual researchers, but also entire research groups or even national research infrastructure providers.



GitHub stars The GitHub project allows users to *star* their favourite projects. This allows users to find them more easily. It also is an indicator of a project's popularity. The chart to the left shows the number of stars that INCEpTION has received over time[29]. To put this into perspective, we also included the number of stars that the older WebAnno annotation editor has gained.

Installation base Upon the first start of the application, INCEpTION asks users if they agree to supply anonymous usage data to a telemetry collection system, which we have operated at TU Darmstadt since August 2019. If the user agrees, the instance sends a regular *alive* signal while it is running, information about which version is used, and a few other details which are fully transparent in the INCEpTION UI. As of Aug 2022, ≈**350 active installations per month** from all over the world currently send us their anonymous usage data (Figure 1). We do not know what percentage of users do not agree to supplying the usage data – different sources on the internet report that it can vary wildly. Piwik, a supplier of telemetry analytics software, suggests it may range from 30%-70% depending on industry.[30] Assuming these opt-in rates, the true number of active installations **may be anywhere from ≈500 to ≈1100**.

INCEpTION is mostly used by individuals and small working groups. However, there are also prominently large installations. Particularly noteworthy is the INCEpTION instance offered by CLARIN-EL to the Greek researcher community.[31]. As of Aug 2022, the largest INCEpTION installation submitting usage data reported a number of ≈**700 enabled accounts**. The highest reported number of users actively logged in and using a single INCEpTION instance at the same time was **35 simultaneous users**. From Jan to Aug 2022, most usage data has been collected from users in **Germany (≈23%)** and in the **U.S.A. (≈20%)**, followed by China, Ireland and France (each ≈6-7%). Overall, usage data from users in **74 countries** has been received in this time frame.

[26] https://github.com/inception-project
[27] https://github.com/dkpro/dkpro-cassis
[28] https://github.com/inception-project/inception/issues
[29] https://seladb.github.io/StarTrack-js/#/preload?r=webanno,webanno&r=inception-project,inception
[30] https://piwik.pro/blog/how-to-do-useful-analytics-without-personal-data/
[31] https://inventory.clarin.gr/tool-service/844

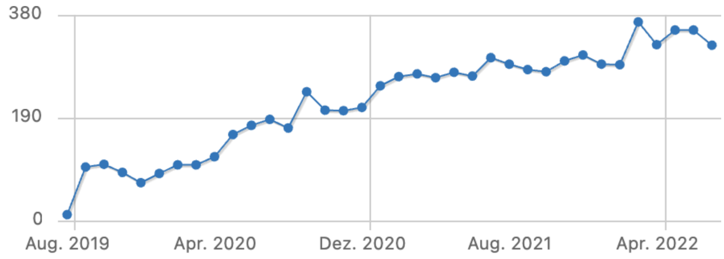


Figure 1: INCEpTION installations per month reporting anonymous usage data over time.

Featured use-cases During the project, we have been in contact with various (groups of) users that were either seeking to use INCEpTION, were already using it, or had been using it before contact with our group was established. From a small portion of these, we have collected use-case descriptions and feature them on the INCEpTION website[32]. As of Oct 2022, the use-case gallery includes 19 use-case descriptions from a range of different application domains:

- **Annotation and Analysis of Moralization Practices**[33]; Maria Becker, Germanistisches Seminar, Universität Heidelberg, Germany;
- **Automated Analysis of Media Suicide Reporting**[34]; by Markus Schäfer, Communication Science Research Group, Johannes Gutenberg-Universität, Mainz, Germany;
- **Cross-Document Event Coreference**[35] – Annotating mentions of events in news articles and linking them to an event in a knowledge base; by Michael Bugert, UKP Lab, Technical University of Darmstadt, Germany;
- **Cited Loci**[36] – Annotating references to classical authors and their works; by Matteo Romanello, Digital Humanities Laboratory, EPFL, Switzerland;
- **Company Inventory**[37] – Efficient annotation of company inventory data; by Maria Biryukov, University of Luxembourg, Center for Contemporary and Digital History;
- **Digital Athenaeus**[38] – Named entity annotation for ancient Greek; by Monica Berti, Department of Digital Humanities, Universität Leipzig, Germany;
- **FAMULUS**[39] – Annotate student essays with categories of diagnostic reasoning; by Claudia Schulz, UKP Lab, Technische Universität Darmstadt , Germany;
- **GGPONC**[40] – German guideline program in oncology NLP corpus; by Florian Borchert, Hasso-Plattner-Institut Digital Health Center, Universität Potstam, Germany;
- **The interactional language of Andreas Gryphius**[41]; by Marcel Fladrich, Linguistik des Deutschen / Medienlinguistik, Universität Hamburg, Germany;
- **GUM Corpus**[42] – Georgetown University Multilayer Corpus; by Amir Zeldes, Department of Linguistics, Georgetown University, Washington, D.C., USA;
- **Impresso**[43] – Critical text mining of newspaper archives; by Matteo Romanello, Digital Humanities Laboratory, EPFL, Switzerland;
- **Legal Argument Mining**[44] – Annotating arguments in the judgements of the European Court of Human Rights; by Ivan Habernal, Trustworthy Human Technologies, Technical University of Darmstadt, Germany;

[32] https://inception-project.github.io/use-cases/
[33] https://inception-project.github.io/use-cases/aamp/
[34] https://inception-project.github.io/use-cases/asmsr/
[35] https://inception-project.github.io/use-cases/cdcr/
[36] https://inception-project.github.io/use-cases/cited-loci/
[37] https://inception-project.github.io/use-cases/company-inventory/
[38] https://inception-project.github.io/use-cases/digital-athenaeus/
[39] https://inception-project.github.io/use-cases/famulus/
[40] https://inception-project.github.io/use-cases/ggponc/
[41] https://inception-project.github.io/use-cases/gryphius/
[42] https://inception-project.github.io/use-cases/gum/
[43] https://inception-project.github.io//use-case-gallery/impresso/
[44] https://inception-project.github.io/use-cases/legal-argument-mining/

- **Multilingual Anonymisation for Public Administrations (MAPA)**[45] – Named entity annotation for anonymisation.; by Victoria Arranz, Evaluations and Language resources Distribution Agency (ELDA), Paris, France;
- **Mining and Modeling Text (MiMoText)**[46] – Annotating scholarly works and linking them to a knowledge base; by Tinghui Duan, Maria Hinzmann, Anne Klee, Johanna Konstanciak, Julia Röttgermann, Christof Schöch and Moritz Steffes, Univ. of Trier, Trier Center for Digital Humanities;
- **Portuguese Parish Memories (1758-1761)**[47] – Annotating language variation and named entities; by Helena Freire Cameron, Fernanda Olival and Renata Vieira from CIDEHUS, University of Évora and VALORIZA, Polytechnics of Portalegre, Portugal;
- **PO-EMO**[48] – Annotating poetry for its emotional impact; by Steffen Eger, Independent Research Group Leader, Department of Computer Science, Technische Universität Darmstadt, Germany;
- **Rapid adaptation of event detection**[49] – Using external search to select and annotate training examples; by Ryan Gabbard and Marjorie Freedman, VISTA group at the Information Sciences Institute at the University of Southern California, USA;
- **SOFC-Exp Corpus**[50] – Annotating experiments in scientific publications on solid oxide fuel cells; by Annemarie Friedrich and Heike Adel, Bosch Center for Artificial Intelligence, Renningen, Germany;
- **A World of Possibilities**[51] – Semantic annotation of modality in a diachronic Latin corpus; by Helena Bermúdez Sabel, Université de Neuchâtel, Switzerland.

Invited talks We have been invited to give talks on INCEpTION to a broad range of different audiences.

- (02. Feb 2017) **INCEpTION - Towards a community-oriented annotation platform with assistive features** to the research group of Jonas Kuhn at the IMS Stuttgart. Presenter: Richard Eckart de Castilho. Invited by: Jonas Kuhn.
- (08. May 2017) **Towards an Infrastructure for the Distributed Exploration and Annotation of Large Corpora and Knowledge Bases** [52] as part of the FEAST Talk Series at the Universität des Saarlandes. Presenter: Richard Eckart de Castilho. Invited by: Elke Teich (Universität des Saarlandes).
- (12.-13. Jul 2017) **Web-based Annotation: from WebAnno to INCEpTION**[53,54]. Historical Text Reuse Data Workshop within the framework of the BMBF funded Global Philology Planning Project, Universität Leipzig. Presenter: Richard Eckart de Castilho. Workshop organized by: Monica Berti (Universität Leipzig) and Gregory R. Crane (Universität Leipzig).
- (11 Dez 2018) **Linking Text and Knowledge using the INCEpTION annotation platform** as part of the colloquium talks series of the Institute of Linguistics and Literary Studies, TU Darmstadt. Presenter: Richard Eckart de Castilho. Invited by: Sabine Bartsch (TU Darmstadt)
- (12.-17. May 2019) **Linking Knowledge and Text using the INCEpTION text annotation platform**[55] at the 3rd Summer Datathon on Linguistic Linked Open Data (SD-LLOD-19) at Dagstuhl, Germany. Presenter: Richard Eckart de Castilho. Workshop organized by Christian Chiarcos (Goethe Universität Frankfurt), John Philip McCrae (Insight Centre for Data Analytics, NUI Galway), Jorge Gracia (University of Zaragoza).
- (31. May-01 Jun 2019) **Collaborative morpho-syntactic annotation in INCEpTION**[56] at the Digital Editions in Practice workshop, Perseus Digital Library, Tufts University, Medford, MA, USA. Presenter: Richard Eckart de Castilho. Workshop by: Gregory Crane (Tufts University), Lisa Cerrato (Perseus Digital Library) and others.

[45] https://inception-project.github.io/use-cases/mapa/

[46] https://inception-project.github.io/use-cases/mimotext/

[47] https://inception-project.github.io/use-cases/parish-memories/

[48] https://inception-project.github.io/use-cases/po-emo/

[49] https://inception-project.github.io/use-cases/rapid-events/

[50] https://inception-project.github.io/use-cases/sofc/

[51] https://inception-project.github.io/use-cases/woposs/

[52] http://feast.coli.uni-saarland.de

[53] https://web.archive.org/web/20170716092256/http://www.dh.uni-leipzig.de/wo/historical-text-reuse-data-workshop/

[54] https://web.archive.org/web/20210424060848/http://www.dh.uni-leipzig.de/wo/wp-content/uploads/2017/05/Historical-Text-Reuse-Data-Workshop-Report.pdf

[55] https://datathon2019.linguistic-lod.org

[56] http://sites.tufts.edu/perseusupdates/2019/01/31/digital-editions-in-practice-a-two-day-workshop/

- (13.-15. Nov 2019) **Immersion Annotation with INCEpTION**[57] in the workshop *Mining Goodreads. A text similarity-based approach to measure reader absorption*, Università di Verona, Italy. Presenter: Ute Winchenbach; Workshop organized by: Simone Rebora (University of Verona, University of Basel), Piroska Lendvai (University of Basel), and Moniek Kuijpers (University of Basel).
- (15. Nov 2019) **DKPro Core and INCEpTION - Modular, interoperable, reusable TDM tools for the community**[58] at the Visa TM Day at the Ministry of Higher Education, Research and Innovation, Paris, France. Presenter: Richard Eckart de Castiilho. Organizer: Claire Nedellec (INRA) et al.
- (12. Feb 2019) **Linking Biomedical Publications and Knowledge using the INCEpTION annotation platform**[59] at the Biomedical Linked Annotation Hackathon (BLAH) 5 organized by the Database Center for LifeScience (DBCLS) in Kashiwa, Japan. Presenter: Richard Eckart de Castilho. Invited by: Jin-Dong Kim (DBCLS, ROIS)
- (23 Jul 2021) **INCEpTION: Durch Annotation zur effizienten digitalen Textanalyse**[60] in the context of the workshop *Einführung in Digital Humanities* at the Westfälische Wilhelms-Universität Münster. Presenter: Ute Winchenbach. Workshop organized by: Anna Gordon (WWU Münster).

Invited workshops and tutorials We have also been invited several times during the project to give workshops and tutorials on INCEpTION.

- (24 Apr & 03 Jun 2020) **INCEpTION-Workshop**[61] in the context of the MiMoText project, Trier Center for Digital Humanities, Universität Trier. Presenter: Anna-Felicitas Hausmann. Invited by: Christoph Schöch (Universität Trier) and Maria Hinzmann (Universität Trier).
- (09.-10. Jun 2021, half-day each) **Introduction to the semantic annotation platform INCEpTION**[62] in the context of the PhD program in *Applied Linguistics: Argumentation in Professional Practice* at ZHAW, Zürich, Switzerland. Presenter: Ute Winchenbach. Invited by: Cerstin Mahlow (ZHAW) and Eva Kuske (ZHAW).
- (19. May 2020) **INCEpTION: Efficient text annotation using human-in-the-loop technology**[63] as part of the seminar series at the Center of Modeling, Simulation and Interactions of the Université Côte d'Azur. Presenter: Ute Winchenbach. Invited by: Marco Corneli (Université Côte d'Azur).
- (29.-30. Sep 2020, 120min each) **INCEpTION** tutorial to members of the CIDEHUS-UÉ group at the University of Évora and their collaborators at University of Portalegre in Portugal and at USP and UFSM in Brazil. Presenter: Anna-Felicitas Hausmann. Invited by: Ivo Santos (University of Évora) and Fernanda Olival (University of Évora)
- (06 Jul 2021, 90min) **The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation**[64] as part of the GSCL Research and Tutorial Talks series. Presenter: Richard Eckart de Castilho. Invited by: Annemarie Friedrich (GSCL).
- (24.-25. Mar 2022, half-day each) Part of the PhD seminar **Annotation and Modeling of Textual Data: Concepts and Tools**[65]. In the context of *Doctoral Programme in Applied Linguistics: Managing Languages, Arguments and Narratives in the Datafied Society* at ZHAW, Zürich, Switzerland. Presenter: Irina Bigoulaeva. Invited by: Cerstin Mahlow (ZHAW) and Maren Runte (ZHAW).

Third-party workshops and other events The following events related to INCEpTION were organized by third parties. We were only minimally involved. These events underline the wider impact of the project.

- (21.-25. Jun 2021, 10.-14. Jan 2022) **New Languages for NLP Building Linguistic Diversity in the Digital Humanities – Workshops I and II**[66]. INCEpTION was introduced as the manual annotation tool for the workshop, which included hands-on sessions. By chance, we discovered the workshop announcement a few days early and, since it was a remote workshop, we spontaneously asked if we could listen in and help answer questions from the workshop participants. This was

[57] https://www.univr.it/it/iniziative/-/evento/8820?p_auth=Ec3AOdAd

[58] https://journees.inra.fr/visa-tm-day/

[59] https://www.youtube.com/watch?v=CN4_dPGVQns

[60] https://www.uni-muenster.de/imperia/md/content/mittellatein/aushang_digital_humanities_aktuell.pdf

[61] https://mimotext.uni-trier.de/activities

[62] https://www.zhaw.ch/storage/linguistik/studium/doktoratsprogramm/workshop-inception.pdf

[63] https://twitter.com/MSI_UCA/status/1255806213935304705?s=20&t=urqPBJu4n5RFXwKQoClD5A

[64] https://gscl.org/2021/07/06/annemariefriedrich-2.html

[65] https://www.zhaw.ch/storage/linguistik/studium/doktoratsprogramm/schedule-spring-semester-2022.pdf

[66] https://newnlp.princeton.edu/Workshop-I/

happily accepted by the organizers. INCEpTION was also featured in the second workshop[67], although we did not join that one.

- (2020) **HIPE (Identifying Historical People, Places and other Entities)**[68] is a evaluation campaign on named entity processing on historical newspapers in French, German and English, which was organized in the context of the impresso project and run as a CLEF 2020 Evaluation Lab. INCEpTION was used to prepare the gold standard data for the shared task. Matteo Romanello, one of the organizers of the shared task, was a guest at our INCEpTION kick-off workshop. Richard Eckart de Castilho was also asked to serve on the shared task's advisory board.

Select related third-party blog posts, videos and talks (last accessed 16. Aug 2022) These blog posts, videos and talks related to INCEpTION were authored by third parties and illustrate the wider impact of the project. Unless stated otherwise, we were not involved in any way.

- **Collecting annotations from British Library staff**[69] by Daniel van Strien, British Library (posted 23. Sep 2019).
- **Named Entity Annotation for Ancient Greek with INCEpTION**[70] by Monica Berti, University of Leipzig (posted 07. Feb 2020). We maintain loose contact with Monica Berti since she invited us for a presentation on INCEpTION a text re-use workshop (cf. invited talks section) in Leipzig in 2017.
- **INCEpTION: A Semantic Annotation Platform**[71] by Shohreh Haddadan, Digital History and Hermeneutics DTU, Luxembourg Centre for Contemporary and Digital History, University of Luxembourg (posted 13. Feb 2020)
- **The best free labeling tools for text annotation in NLP**[72] by Fabian Gringel, dida Datenschmiede GmbH, Berlin (posted 30. Mar 2020)
- **INCEpTION**[73] by Mareike Schumacher and Kristina Becker, forTEXT Projekt, Darmstadt. The authors contacted us with a questionaire on certain aspects of the INCEpTION software which we provided back to them. (posted 05. Apr 2021)
- **As You Like It: Event Annotation with INCEpTION**[74] by Laska Laskova, Kiril Simov, Petya Osenova, Iva Anastasova, Preslava Georgieva, IICT-BAS Bulgaria at CLARIN CAC 2020.

Related 3rd party open source software projects These open source projects related to INCEpTION were initiated by 3rd parties and indicate the wider impact of the project. We provided advice to them.

- **INCEpTALYTICS**[75] – an easy-to-use API for analyzing INCEpTION annotation projects.
- **CORLINCEpTION converter**[76] – aims at helping compatibility of corpora with INCEpTION by converting standard XML corpora (e.g. TEI-XML) to UIMA.
- **PyCaprio**[77] – a Python client to the INCEpTION annotation tool.

[67] https://newnlp.princeton.edu/Workshop-II/
[68] https://impresso.github.io/CLEF-HIPE-2020/
[69] https://livingwithmachines.ac.uk/collecting-annotations-from-british-library-staff
[70] https://videolectures.net/clarinannualconference2019_berti_named_entity/
[71] https://dhh.uni.lu/2020/02/13/inception-a-semantic-annotation-platform/
[72] https://dida.do/blog/the-best-free-labeling-tools-for-text-annotation-in-nlp
[73] https://fortext.net/tools/tools/inception
[74] https://clada-bg.eu/images/news/Laska_INCEPTION_PRES.pdf
[75] https://github.com/ltl-ude/inceptalytics
[76] https://github.com/Consortium-CORLI/converter_for_inception
[77] https://github.com/JavierLuna/pycaprio

3 Summary

Short, plain-language presentation of the work performed, the progress made, and the project results achieved The project has developed INCEpTION, an open source text annotation platform software. It has become a popular go-to solution for users with the need to annotate text across a wide range of disciplines and use-cases[78].

The INCEpTION annotation platform incorporates three functional pillars: automatic annotation suggestions using machine learning, knowledge management and search. These support in particular tasks such as named entity annotation and named entity linking, but as the use-cases mentioned above highlight, the platform is by far not limited to such tasks. INCEpTION has evolved to fully replace the earlier WebAnno tool, while still remaining largely backwards compatible and offering WebAnno users and easy upgrade path. With respect to automatic annotation suggestions, INCEpTION not only brings several machine learning algorithms out of the box, but also allows to connect custom external machine learning algorithms or even text analysis services such as European Language Grid and CLARIN WebLicht. With respect to knowledge management, INCEpTION supports the widely used SPARQL protocol which allows connecting to many knowledge providers such as offered by the ZBW Leibnitz Information Centre for Economics, DBPedia, Wikidata, the Food and Agriculture Organization (FAO) of the United Nations and many more.

In Aug 2022, ≈350 active installations in 74 countries were sending us anonymous usage data, with the majority of installations in Germany, the U.S.A., China and France. Considering that up to 70% of the installations may have decided to opt out of anonymous data submissions, the actual number of installations may be over 1 000. Over 430 people have *starred* the INCEpTION project on GitHub. Over 200 people have subscribed to the project's mailing list and generated over 1 300 messages in over 400 conversations. Until Aug 2022, over 160 users have opened over 400 issues with bug reports, feature requests and questions on our GitHub issue tracker, contributing to the total of over 1 900 issues tracked there at the time (1 700 of which have been resolved).

Necessary deviations from the original project plan Based on the interaction with the user community during the project, we deviated in various aspects from the original ideas presented in the project proposal. The most significant deviation was the switch from focusing compute clusters to process data to being interoperable with GPU-based machine learning. The advent of GPU-enabled deep learning caused to community to strongly focus on this technology and artificial neural networks have achieved amazing new results over the last years. It became clear early in the project that we needed to support these new technologies and so we decided to also switch our focus from cluster-based processing to interoperability with GPU-based machine learning, in particular the ability to interface with popular Python-based packages. The architectural changes involved in this switch also enabled us to interoperate with external text analysis services such as those offered by the European Language Grid, CLARIN WebLicht, LAPPS Grid (deprecated) or Huggingface (experimental).

4 Further Work and Plans

What activities building on the results achieved are planned, if any? The highly modularized INCEpTION annotation platform code base offers a strong basis for future development. Despite the broad range of features and the generic nature of the platform, there are still many use-cases and workflows which the platform does not support well, e.g. the ability for multiple users to view a document simultaneously and to collaborate on it in such a manner that every user immediately sees what the other users do. We are looking towards incorporating new technologies to enable such use-cases. Another means of enabling new use-cases would be the ability to easily built custom annotation editor interfaces that can be used within the platform. Therefore, we are currently working on developing new programming interfaces and protocols to enable such custom interfaces.

[78] https://inception-project.github.io/use-cases/

References

[1] B. Boullosa, R. E. de Castilho, A. Geyken, L. Lemnitzer, and I. Gurevych. A tool for extracting sense-disambiguated example sentences through user feedback. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, number TUD-CS-2017-0039, pages 69–72. Association for Computational Linguistics, April 2017.

[2] B. Boullosa, R. E. de Castilho, N. Kumar, J.-C. Klie, and I. Gurevych. Integrating Knowledge-Supported Search into the INCEpTION Annotation Platform. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing - Demo Papers*, pages 127–132, Aug. 2018.

[3] R. E. de Castilho, N. Ide, J.-D. Kim, J.-C. Klie, and K. Suderman. A multi-platform annotation ecosystem for domain adaptation. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 189–194, Florence, Italy, August 2019. Association for Computational Linguistics. Veranstaltungstitel: 13th Linguistic Annotation Workshop.

[4] R. E. de Castilho, N. Ide, J.-D. Kim, J.-C. Klie, and K. Suderman. Towards cross-platform interoperability for machine-assisted annotation. *Genomics & Informatics*, 17(2):e19., Juni 2019.

[5] R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, and I. Gurevych. INCEpTION - Corpus-based Data Science from Scratch. In *Digital Infrastructures for Research (DI4R) 2018*, Oct. 2018.

[6] R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, and I. Gurevych. Linking Text and Knowledge using the INCEpTION annotation platform. In *Proceedings of the 14th eScience IEEE International Conference*, pages 327–328, Oct. 2018.

[7] J.-C. Klie. INCEpTION: Interactive Machine-assisted Annotation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, pages 105–105, July 2018.

[8] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018.

[9] J.-C. Klie, R. E. de Castilho, and I. Gurevych. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. pages 6982–6993, Juli 2020. Veranstaltungstitel: The 58th annual meeting of the Association for Computational Linguistics (ACL 2020).

[10] J.-U. Lee, J.-C. Klie, and I. Gurevych. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373, Juni 2022.

[11] C. Scheunemann, J. Naumann, M. Eichler, K. Stowe, and I. Gurevych. Data collection and annotation pipeline for social good projects. November 2020. Veranstaltungstitel: AI for Social Good - AAAI Fall Symposium 2020.

[12] M. Wu, N. S. Moosavi, A. Rücklé, and I. Gurevych. Improving QA generalization by concurrent modeling of multiple biases. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 839–853. Association for Computational Linguistics, Oktober 2020.