Constructing our own data warehouse will be critical to the success of our project. We will develop our own Extract, Transform, Load (ETL) procedure (Kimball & Caserta, 2004) to stream the flow of data directly into a carefully designed relational database.

The extraction will require to download the available datasets (RePORT, NSF award, PubMed Database, UMETRICS, Clinicaltrials.gov, STATT), and to convert them from `XML` or `CSV` to tables in a relational database (`SQL`). High-quality extraction and conversion will be performed using state-of-the-art MySQL's `XML` functions, Microsoft's `XML` Bulk Load or pureXML's built-in shredding capacities (Lee, Mani, & Chu, 2002, 2003; Nicola & Kumar-Chatterjee, 2009). Each dataset comes with a precise, documented schema, that will be leveraged, if needed using genetic algorithms (Ng, Kong, & Chan, 2004), manual annotations (Sunchu, 2016), or semantics constraints (Lv & Yan, 2006) to insure the best possible quality of the extracted data.

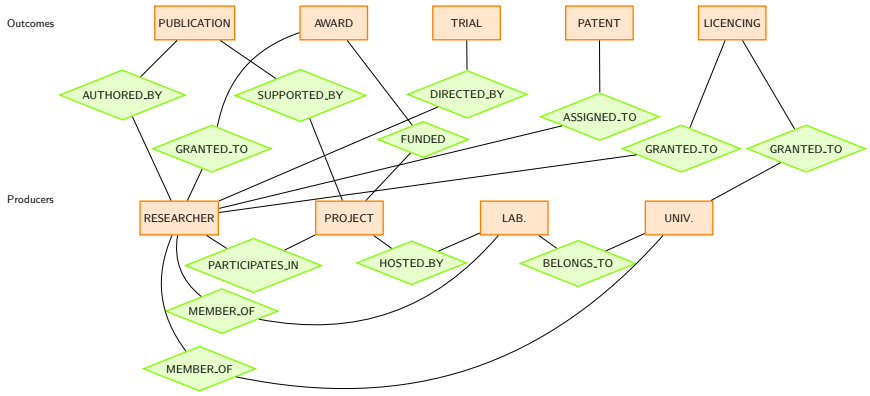
The transformation will mainly consist in identifying records, or data sets, representing the same real-world entity (e.g., author, lab, university). This extremely well-known problem, called record linkage, entity resolution or matching, is extensively studied (Elmagarmid, Ipeirotis, & Verykios, 2007; Winkler, 1999, 2006), but numerous challenges remain. For instance, the absence of a best match algorithm in all generality (Köpcke & Rahm, 2010) will require from us to experiment, to get an effective and efficient combination of different techniques. Preliminary study of the datasets suggests that:

- "Blocking strategies" (Lehti & Fankhauser, 2005; Rohan Baxter & Churches, 2003; Wang, 2016) will be needed to reduce the search space.
- Since direct identifiers are available and of good quality, deterministic methods should have good results (Dusetzina et al., 2014; Howe, Lake, & Shen, 2006; On, Lee, Choi, & Park, 2014), and efficient matching algorithm (Benjelloun et al., 2009) should give good results.
- Overlaps and pre-existing linkages in the datasets will give us "hints" that can be exploited to find matching records (Whang, Marmaros, & Garcia-Molina, 2013).
- We will be able to use linkage at multiple levels, and to benefits from the technologies of group linkage (On, Koudas, Lee, & Srivastava, 2007).

Even if no privacy issues are foreseen, should they arise, then methods to link datasets without disclosing sensitive information will be used (Karapiperis, Verykios, Katsiri, & Delis, 2016; Kum, Krishnamurthy, Machanavajjhala, Reiter, & Ahalt, 2013; Kum et al., 2019; Vatsalan, Christen, & Verykios, 2013; Vidanage, Ranbaduge, Christen, & Schnell, 2019). The latest development in benchmarking will be used to asses the quality of the linkage (Ferrante & Boyd, 2012; Hand & Christen, 2018; Köpcke & Rahm, 2010), but our ultimate unit of measure will be our goal (Adil, Tengku Izhar, Torabi, & Bhatti, 2017).

Indeed, the linkage will be directed toward the filling of a carefully crafted entity–relationship model that will be designed by all the team members. The

"evidence" entities (publications, awards, etc.) will be in relationship with our "producer" entities (PIs, projects, etc.) through numerous relationships that will allow us to model a large variety of situations.



Attributes are not represented for concision. Every relationship has a cardinality ratio of $M : N$ and no participation constraint.

Loading the extracted, transformed data will require to preserve backtracking possibilities: the origin of the data and the reason why entities were matched needs to be precisely documented, so that corrections can be applied when new datasets or better matching algorithms become available. Weekly and monthly updates of the datasets will be propagated to our warehouse using similar or improved techniques, and will enrich our basis for analysis with almost real-time information. Cohorts of principal investigators will then be extracted and updated from this data warehouse using subgraph patterns (Moustafa, Kimmig, Deshpande, & Getoor, 2014) and techniques to identify cliques in large networks (Conte, Grossi, & Marino, 2019; Rossi, Gleich, Gebremedhin, & Patwary, 2014)

Adil, T., Tengku Izhar, T. A., Torabi, T., & Bhatti, M. (2017). Record linkage in organisations: A review and directions for future research. *International Journal of Data Science*, *2*, 325–351. https://doi.org/10.1504/IJDS.2017.088103

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: A generic approach to entity resolution. *The VLDB Journal*, *18*(1), 255–276. https://doi.org/10.1007/s00778-008-0098-x

Conte, A., Grossi, R., & Marino, A. (2019). Large-scale clique cover of real-world networks. *Information and Computation*, 104464. https://doi.org/10.1016/j.ic.2019.104464

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., & Carpenter, W. R. (2014). *Linking Data for Health Services Research: A Framework and Instructional Guide*. Retrieved from Agency for Healthcare Research; Quality (US) website: https://www.ncbi.nlm.nih.gov/books/NBK253313/pdf/Bookshelf_NBK253313.pdf

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record

detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16. https://doi.org/10.1109/TKDE.2007.250581

Ferrante, A., & Boyd, J. (2012). A transparent and transportable methodology for evaluating data linkage software. *Journal of Biomedical Informatics*, *45*(1), 165–172. https://doi.org/10.1016/j.jbi.2011.10.006

Hand, D., & Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, *28*(3), 539–547. https://doi.org/10.1007/s11222-017-9746-6

Howe, H. L., Lake, A. J., & Shen, T. (2006). Method to assess identifiability in electronic data files. *American Journal of Epidemiology*, *165*(5), 597–601. https://doi.org/10.1093/aje/kwk049

Karapiperis, D., Verykios, V. S., Katsiri, E., & Delis, A. (2016). A tutorial on blocking methods for privacy-preserving record linkage. In I. Karydis, S. Sioutas, P. Triantafillou, & D. Tsoumakos (Eds.), *Algorithmic aspects of cloud computing* (pp. 3–15). Cham: Springer International Publishing.

Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data.* Wiley.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*(2), 197–210. https://doi.org/10.1016/j.datak.2009.10.003

Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M. K., & Ahalt, S. (2013). Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association*, *21*(2), 212–220. https://doi.org/10.1136/amiajnl-2013-002165

Kum, H.-C., Ragan, E. D., Ilangovan, G., Ramezani, M., Li, Q., & Schmit, C. (2019). Enhancing privacy through an interactive on-demand incremental information disclosure interface: Applying privacy-by-design to record linkage. *Fifteenth symposium on usable privacy and security (SOUPS 2019)*. Retrieved from https://www.usenix.org/conference/soups2019/presentation/kum

Lee, D., Mani, M., & Chu, W. W. (2002). Effective schema conversions between XML and relational models. *In european conference on artificial intelligence (ECAI)*, 3–11.

Lee, D., Mani, M., & Chu, W. W. (2003). Schema conversion methods between XML and relational models. In B. Omelayenko & M. C. A. Klein (Eds.), *Knowledge transformation for the semantic web* (pp. 1–17). IOS Press.

Lehti, P., & Fankhauser, P. (2005). A precise blocking method for record linkage. In A. M. Tjoa & J. Trujillo (Eds.), *Data warehousing and knowledge discovery* (pp. 210–220). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lv, T., & Yan, P. (2006). Mapping DTDs to relational schemas with semantic constraints. *Information and Software Technology*, *48*(4), 245–252. https://doi.org/10.1016/j.infsof.2005.05.001

Moustafa, W. E., Kimmig, A., Deshpande, A., & Getoor, L. (2014). Subgraph pattern matching over uncertain graphs with identity linkage uncertainty. *2014 IEEE 30th international conference on data engineering*, 904–915. https://doi.org/10.1109/ICDE.2014.6816710

Ng, V., Kong, C. C., & Chan, S. (2004). Mapping XML schema to relations

using genetic algorithm. In M. Gh. Negoita, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based intelligent information and engineering systems* (pp. 232–245). https://doi.org/10.1007/978-3-540-30134-9_33

Nicola, M., & Kumar-Chatterjee, P. (2009). *DB2 pureXML cookbook: Master the power of the IBM hybrid data server*. IBM Press.

On, B.-W., Koudas, N., Lee, D., & Srivastava, D. (2007). Group linkage. In R. Chirkova, A. Dogac, M. T. Özsu, & T. K. Sellis (Eds.), *Proceedings of the 23rd international conference on data engineering, ICDE 2007, the marmara hotel, istanbul, turkey, april 15-20, 2007* (pp. 496–505). https://doi.org/10.1109/ICDE.2007.367895

On, B.-W., Lee, I., Choi, G. S., & Park, H.-S. (2014). Discriminative and deterministic approaches towards entity resolution. *Journal of Intelligent Information Systems*, *43*(1), 101–127. https://doi.org/10.1007/s10844-014-0308-5

Rohan Baxter, P. C., & Churches, T. (2003). A comparison of fast blocking methods for record linkage. *Proceedings of the workshop on data cleaning, record linkage and object consolidation at the ninth ACM SIGKDD international conference on knowledge discovery and data mining, washington DC*, 25–27. Retrieved from http://users.cecs.anu.edu.au/~christen/publications/kdd03-6pages.pdf

Rossi, R. A., Gleich, D. F., Gebremedhin, A. H., & Patwary, Md. M. A. (2014). Fast maximum clique algorithms for large graphs. *Proceedings of the 23rd international conference on world wide web*, 365–366. https://doi.org/10.1145/2567948.2577283

Sunchu, V. K. (2016). *A flexible schema-aware mapping of XML data into relational models* (Master's thesis, The University of Oklahoma, College of Engineering, School of Computer Science). Retrieved from https://hdl.handle.net/11244/44903

Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, *38*(6), 946–969. https://doi.org/10.1016/j.is.2012.11.005

Vidanage, A., Ranbaduge, T., Christen, P., & Schnell, R. (2019). Efficient pattern mining based cryptanalysis for privacy-preserving record linkage. *35th IEEE international conference on data engineering, ICDE 2019, macao, china, april 8-11, 2019*, 1698–1701. https://doi.org/10.1109/ICDE.2019.00176

Wang, P. (2016). Blocking strategies for performing entity resolution in a distributed computing environment (PhD thesis, Graduate School, University of Arkansas at Little Rock; p. 102). Retrieved from https://search.proquest.com/docview/1883385264?accountid=12365

Whang, S. E., Marmaros, D., & Garcia-Molina, H. (2013). Pay-as-you-go entity resolution. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 1111–1124. https://doi.org/10.1109/TKDE.2012.43

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Statistical Research Division, U.S. Census Bureau.

Winkler, W. E. (2006). *Overview of record linkage and current research directions*. Statistical Research Division, U.S. Census Bureau.