



Universiteit
Leiden
The Netherlands

Design and development of a comprehensive data management platform for cytomics: cytomicsDB

Larios Vargas, E.

Citation

Larios Vargas, E. (2015, November 25). *Design and development of a comprehensive data management platform for cytomics: cytomicsDB*. Retrieved from <https://hdl.handle.net/1887/36426>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/36426>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/36426> holds various files of this Leiden University dissertation

Author: Larios Vargas, Enrique

Title: Design and development of a comprehensive data management platform for cytomics : cytomicsDB

Issue Date: 2015-11-25

Bibliography

- [Alv] 96 well plate format, <http://reinnervate.com/why-alvetex/alvetex-product-catalogue/>, Accessed: 2015-09-25. (cited on page 11).
- [AMF00] George Avery, Charles McGee, and Stan Falk, *Product review: Implementing lims: A "how-to" guide.*, *Analytical Chemistry* **72** (2000), no. 1, 57 A–62 A. (cited on page 20).
- [Ber03] Bryan Bergeron, *Bioinformatics computing*, Prentice Hall/Professional Technical Reference, Upper Saddle River, NJ, 2003. (cited on page 82).
- [BKMC⁺12] D.A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers, *Genbank*, *Nucleic Acids Research* **40** (2012), no. Database issue, D48–D53. (cited on page 63).
- [BLW⁺08] E. Bruford, M. Lush, M. Wright, T. Sneddon, S. Povey, and E. Birney, *The hgnc database in 2008: a resource for the human genome.*, *Nucleic acids research* **36** (2008), no. Database issue. (cited on page 60).
- [BMK09] Peter A. Boncz, Stefan Manegold, and Martin L. Kersten, *Database architecture evolution: Mammals flourished long before dinosaurs became extinct*, *PVLDB* **2** (2009), no. 2, 1648–1653. (cited on pages 34 and 71).
- [BN06] J. Bleiholder and F. Naumann, *Conflict handling strategies in an integrated information system.* (cited on pages 57, 58 and 61).
- [BN09] ———, *Data fusion*, *ACM Comput. Surv.* **41** (2009), no. 1, 1:1–1:41. (cited on pages 57, 61 and 72).
- [Bon02] Peter A. Boncz, *Monet: A next-generation dbms kernel for query-intensive applications*, Ph.d. thesis, Universiteit van Amsterdam, Amsterdam, The Netherlands, May 2002. (cited on pages 33, 34, 71 and 85).

- [BSV11] L. M. F. Bertens, J. Slob, and F.J. Verbeek, *A generic organ based ontology system, applied to vertebrate heart anatomy, development and physiology.*, J. Integrative Bioinformatics **8** (2011), no. 2. (cited on page 52).
- [BV08] M. Belmamoune and F. J. Verbeek, *Data integration for spatio-temporal patterns of gene expression of zebrafish development: the gems database*, Journal of Integrative Bioinformatics **5(2):92** (2008). (cited on page 39).
- [BWTB08] E. Bolton, Y. Wang, P. Thiessen, and S. Bryant, *Chapter 12 - pubchem: Integrated platform of small molecules and biological activities*, Annual Reports in Computational Chemistry, vol. 4, Elsevier, 2008, pp. 217 – 241. (cited on page 58).
- [CDV08] David Chen, Guy Doumeingts, and François Vernadat, *Architectures for enterprise integration and interoperability: Past, present and future*, Comput. Ind. **59** (2008), no. 7, 647–659. (cited on page 17).
- [CFK⁺11] C. Colmsee, Steffen Flemming, M. Klapperstuck, M. Lange, and U. Scholz, *A case study for efficient management of high throughput primary lab data*, BMC Research Notes **4** (2011), no. 1, 413. (cited on page 50).
- [CJL⁺06] A. Carpenter, T. Jones, M. Lamprecht, C. Clarke, I. Kang, O. Friman, D. Gertin, J. Chang, R. Lindquist, J. Moffat, P. Golland, and D. Sabatini, *Cellprofiler: image analysis software for identifying and quantifying cell phenotypes*, Genome Biology **7(10)** **10** (2006), no. 7. (cited on page 40).
- [CMS06] T.P.S. Chan, P. Malik, and R. Singh, *An interactive visualization-based approach for high throughput screening information management in drug discovery*, Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, Aug 2006, pp. 5794–5797. (cited on page 51).
- [Cor88] F. Corpet, *Multiple sequence alignment with hierarchical clustering.*, Nucleic Acids Research **16** (1988), no. 22, 10881–10890. (cited on page 74).
- [CYW⁺11] L. Cao, K. Yan, L. Winkel, M. de Graauw, and F.J. Verbeek, *Pattern recognition in high-content cytomics screens for target discovery - case studies in endocytosis*, PRIB 2011, 2011. (cited on pages 28, 29, 30, 47, 48, 50, 81 and 91).
- [DAJ12] S. K. Dubey, A. Anand, and H. Jangala, *Laboratory information and management system: A tool to increase laboratory productivity*, Clinical Research and Regulatory Affairs **29** (2012), no. 2, 46 – 56. (cited on pages 20, 21 and 22).

- [DC08] Ewa Deelman and Ann Chervenak, *Data management challenges of data-intensive scientific workflows*, Proceedings of the 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (Washington, DC, USA), CCGRID '08, IEEE Computer Society, 2008, pp. 687–692. (cited on pages 16 and 17).
- [dGCW⁺13] M. de Graauw, L. Cao, L. Winkel, M. H. A. M. van Miltenburg, S. LeDévédec, M. Klop, K. Yan, C. Pont, V-M. Rogkoti, A. Tijsma, A. Chaudhuri, R. Lalai, L. Price, F. Verbeek, and B. van de Water, *Annexin a2 depletion delays egfr endocytic trafficking via cofilin activation and enhances egfr signaling and metastasis formation*, *Oncogene* (2013). (cited on page 47).
- [DGDLM13] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, *Addressing big data issues in scientific data infrastructure*, Collaboration Technologies and Systems (CTS), 2013 International Conference on, May 2013, pp. 48–55. (cited on page 18).
- [Di13] Z. Di, *Development of automatic image analysis methods for high-throughput and high-content screening*, Ph.d. thesis, Leiden University, Leiden, The Netherlands, December 2013. (cited on page 14).
- [EBG⁺12] Kevin W. Eliceiri, Michael R. Berthold, Ilya G. Goldberg, Luis Ibanez, B. S. Manjunath, Maryann E. Martone, Robert F. Murphy, Hanchuan Peng, Anne L. Plant, Badrinath Roysam, Nico Stuurman, Jason R. Swedlow, Pavel Tomancak, and Anne E. Carpenter, *Biological imaging software tools*, *Nat Meth* **9** (2012), no. 7, 697–710. (cited on pages 11, 12 and 13).
- [FFM05] A. Fuxman, E. Fazli, and R. J. Miller, *Conquer: Efficient management of inconsistent databases*, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (New York, NY, USA), SIGMOD '05, ACM, 2005, pp. 155–166. (cited on page 57).
- [GAK⁺14] Robert Grandl, Ganesh Ananthanarayanan, Srikanth Kandula, Sriram Rao, and Aditya Akella, *Multi-resource packing for cluster schedulers*, Proceedings of the 2014 ACM Conference on SIGCOMM (New York, NY, USA), SIGCOMM '14, ACM, 2014, pp. 455–466. (cited on page 81).
- [GM07] Estelle Glory and Robert F. Murphy, *Automated subcellular location determination and high-throughput microscopy*, *Developmental Cell* **12** (2007), no. 1, 7 – 16. (cited on page 12).
- [GMNR05] M. Gudgin, N. Mendelsohn, M. Nottingham, and H. Ruellan, *Soap message transmission optimization mechanism*, 2005. (cited on page 33).
- [Gro10] The HDF Group, *Hierarchical data format version 5*, 2000-2010. (cited on page 46).

- [GW06] R. C. Gonzalez and R. E. Woods, *Digital image processing (3rd edition)*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006. (cited on page 82).
- [Han02] G. Hannon, *Rna interference*, *Nature* **418** (2002), no. 6894, 244–251. (cited on page 59).
- [Hey02] Stephan Heyse, *Comprehensive analysis of high-throughput screening data*, 2002, pp. 535–547. (cited on page 18).
- [HT03] T. Hey and A. Trefethen, *The data deluge: An e-science perspective*, pp. 809–824, John Wiley Sons, Ltd, 2003. (cited on page 17).
- [JZR⁺08] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis, and T. Madden, *Ncbi blast: a better web interface*, *Nucleic Acids Research* **36** (2008), no. suppl 2, W5–W9. (cited on page 60).
- [KAEM13] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, and William Money, *Big data: Issues and challenges moving forward*, 2014 47th Hawaii International Conference on System Sciences **0** (2013), 995–1004. (cited on page 19).
- [KBK⁺09] A. Kallergi, Y. Bei, P. Kok, J. Dijkstra, J.P. Abrahams, and F.J. Verbeek, *Cyttron: A virtualized microscope supporting image integration and knowledge discovery*, *Cell Death and Disease Series: Proteins Killing Tumour Cells* (2009), 291 – 315. (cited on page 39).
- [KBL⁺08] K. Kohl, G. Basler, A. Ludemann, J. Selbig, and D. Walther, *A plant resource and experiment management system based on the golm plant database as a basic tool for omics research*, *Plant Methods* **4** (2008), no. 1, 11. (cited on page 51).
- [KGG10] S. Kalani, D. Gupta, and R. Gupta, *Good laboratory practices : compliance using lims*, no. 1, 41–43. (cited on page 22).
- [Kos06] Kurt Kosanke, *Iso standards for interoperability: a comparison*, *Interoperability of Enterprise Software and Applications* (Dimitri Konstantas, Jean-Paul Bourrières, Michel Léonard, and Nacer Boudjlida, eds.), Springer London, 2006, pp. 55–64. (cited on page 17).
- [Kri09] Wim P. Krijnen, *Applied statistics for bioinformatics using r*, 2009. (cited on page 86).
- [KS06] M. Keith and M. Schincariol, *Pro ejb 3: Java persistence api (pro)*, Apress, Berkely, CA, USA, 2006. (cited on pages 33 and 85).
- [Lak99] Joseph R. Lakowicz, *Principles of fluorescence spectroscopy*, second ed., Kluwer Academic/Plenum Publishers, 1999. (cited on page 13).

- [LC09] Vebjorn Ljosa and Anne E Carpenter, *Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening*, PLoS Computational Biology **5** (2009), no. 12, e1000603. (cited on page 12).
- [Lev66] V.I. Levenshtein, *Binary codes capable of correcting deletions, insertions and reversals*, Soviet Physics Doklady **10** (1966), 707. (cited on pages 67 and 74).
- [LYdB⁺10] S. LeDévédec, K. Yan, H. de Bont, V. Ghotra, H. Truong, E. Danen, F.J. Verbeek, and B. van de Water, *A systems microscopy approach to understand cancer cell migration and metastasis*, Journal of Cellular and Molecular in Life Science (2010). (cited on pages 28, 30 and 31).
- [LZCV14] E. Larios, Y. Zhang, L. Cao, and F.J. Verbeek, *Cytomicsdb: A metadata-based storage and retrieval approach for high-throughput screening experiments*, Pattern Recognition in Bioinformatics (Matteo Comin, Lukas Käll, Elena Marchiori, Alioune Ngom, and Jagath Rajapakse, eds.), Lecture Notes in Computer Science, vol. 8626, Springer International Publishing, 2014, pp. 72–84. (cited on pages 54 and 85).
- [LZY⁺12] E. Larios, Y. Zhang, K. Yan, Z. Di, S. LeDévédec, F. Groffen, and F.J. Verbeek, *Automation in cytomics: A modern rdbms based platform for image analysis and management in high-throughput screening experiments*, Proceedings of the 1st Int. Conf. on Health Information Science, vol. 7231, 2012, pp. 76–87. (cited on pages 42, 46, 54, 69, 71 and 80).
- [MA04] R. T. Marler and J. S. Arora, *Survey of multi-objective optimization methods for engineering*, Structural and Multidisciplinary Optimization **26** (2004), no. 6, 369–395. (cited on pages 67 and 75).
- [Mah91] R. R. Mahaffey, *Information technology in the laboratory*, Chemometrics and Intelligent Laboratory Systems **13** (1991), no. 1, 69 – 74. (cited on page 20).
- [MAT10] MATLAB, *version 7.10.0 (r2010a)*, The MathWorks Inc., Natick, Massachusetts, 2010. (cited on page 31).
- [MCV⁺06] P. Malik, T. Chan, J. Vandergriff, J. Weisman, J. DeRisi, and R. Singh, *Information management and interaction in high-throughput screening for drug discovery*, Database Modeling in Biology: Practices and Challenges (Z. Ma and J. Chen, eds.), Springer Verlag, 2006. (cited on page 53).
- [MF08] L. Mayr and P. Fuerst, *The future of high-throughput screening*, Journal of Biomolecular Screening (2008). (cited on page 41).

- [Miz02] I. Mizrachi, *Chapter 1 genbank: The nucleotide sequence database*, The NCBI Handbook [Internet] (J. McEntyre and J. Ostell, eds.), Bethesda (MD): National Center for Biotechnology Information (US), 2002. (cited on page 60).
- [MKSP00] D. R. Maglott, K. S. Katz, H. Sicotte, and K. D. Pruitt, *Ncbi's locuslink and refseq*, *Nucleic Acids Research* **28** (2000), no. 1, 126–128. (cited on page 63).
- [Mon] *Boost your data analytics*, <https://www.monetdbolutions.com/solutions/analytics>, Accessed: 2015-02-23. (cited on page 86).
- [Mot99] A. Motro, *Multiplex: A formal model for multidatabases and its implementation*, Next Generation Information Technologies and Systems (RonY. Pinter and Shalom Tsur, eds.), *Lecture Notes in Computer Science*, vol. 1649, Springer Berlin Heidelberg, 1999, pp. 138–158 (English). (cited on page 56).
- [Mot01] ———, *Data integration: Inconsistency detection and resolution based on source properties*, In *Workshop on Foundations of Models for Information Integration (FMII)*, 2001, pp. 429–444. (cited on page 57).
- [NSD⁺10] D. Nix, T. Di Sera, B. Dalley, B. Milash, R. Cundick, K. Quinn, and S. Courdy, *Next generation tools for genomic data generation, distribution, and visualization*, *BMC Bioinformatics* **11** (2010), no. 1, 455. (cited on page 51).
- [NWH⁺10] B. Neumann, T. Walter, J.K. Hériché, J. Bulkescher, H. Erfle, C. Conrad, P. Rogers, I. Poser, M. Held, U. Liebel, C. Cetin, F. Sieckmann, G. Pau, R. Kabbe, A. W'unsche, V. Satagopam, M.H. Schmitz, C. Chappuis, D.W. Gerlich, R. Schneider, R. Eils, W. Huber, J.M. Peters, A.A. Hyman, R. Durbin, R. Pepperkok, and J. Ellenberg, *Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes*, *Nature* **Apr 1** (2010), no. 464(7289), 721–727. (cited on page 31).
- [O'N] Elizabeth J. O'Neil, *Object/relational mapping 2008: Hibernate and the entity data model (edm)*, *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08*. (cited on page 33).
- [PCC13] Pasquale Pagano, Leonardo Candela, and Donatella Castelli, *Data interoperability*, *Data Science Journal* **12** (2013), GRDI19–GRDI25. (cited on page 18).
- [Pok11] Jaroslav Pokorný, *Nosql databases: A step to database scalability in web environment*, *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services (New York, NY, USA), iiWAS '11, ACM, 2011*, pp. 278–283. (cited on page 97).

- [QSY⁺11] Y. Qin, G. Stokman, K. Yan, S. Ramaiahgari, F.J. Verbeek, B. van de Water, and L. Price, *Activation of epac-rap signaling protects against cisplatin-induced apoptosis of mouse renal proximal tubular cells*, *Journal of Biological Chemistry* (2011), In Press. (cited on page 28).
- [Say08] E. Sayers, *E-utilities quick start. entrez programming utilities help [internet]*, <http://www.ncbi.nlm.nih.gov/books/NBK25500/>, 2008, Accessed: 2014 Oct 21. (cited on page 66).
- [SKE07] Paul J. Smith, I.A. Khan, and R.J. Errington, *Cytomics and drug development*, *Cytometry Part A* **71A** (2007), no. 6, 349–351. (cited on pages 14, 16 and 17).
- [SMS⁺02] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, C.J. Stoeckert, and A. Brazma, *Design and implementation of microarray gene expression markup language (mage-ml)*, *Genome Biology* (2002), no. 3(9):RESEARCH0046. (cited on page 39).
- [Sta06] Garrick Staples, *Torque resource manager*, *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing* (New York, NY, USA), SC '06, ACM, 2006. (cited on page 81).
- [SZK⁺11] D.O. Skobelev, T.M. Zaytseva, A.D. Kozlov, V.L. Perepelitsa, and A.S. Makarova, *Laboratory information management systems in the work of the analytic laboratory*, *Measurement Techniques* **53** (2011), no. 10, 1182–1189. (cited on page 20).
- [TGDS04] Kerstin Thurow, Bernd Göde, Uwe Dingerdissen, and Norbert Stoll, *Laboratory information management systems for life science applications*, *Organic Process Research & Development* **8** (2004), no. 6, 970–982. (cited on page 20).
- [Tox] *Visualisation core facilities*, <http://toxicology.leidenuniv.nl/research/facilities>, Accessed: 2015-03-25. (cited on page 12).
- [Whi09] Tom White, *Hadoop: The definitive guide*, 1st ed., O'Reilly Media, Inc., 2009. (cited on page 97).
- [WSP⁺07] M. Wendl, S. Smith, C. Pohl, D. Dooling, A. Chinwalla, K. Crouse, T. Hepler, S. Leong, L. Carmichael, M. Nhan, B. Oberkfell, E. Mardis, L. Hillier, and R. Wilson, *Design and implementation of a generalized laboratory data model*, *BMC Bioinformatics* **8** (2007), no. 1, 362. (cited on pages 42 and 51).

- [Yan13] K. Yan, *Image analysis and platform development for automated phenotyping in cytomics*, Ph.d. thesis, Leiden University, Leiden, The Netherlands, November 2013. (cited on pages 31 and 82).
- [YLL⁺11] K. Yan, E. Larios, S. LeDévédec, B. van de Water, and F.J. Verbeek, *Automation in cytomics: Systematic solution for image analysis and management in high throughput sequences*, Proceedings IEEE Conf. Engineering and Technology (CET 2011), vol. 7, 2011, pp. 195–198. (cited on pages 10, 11, 13, 14, 27, 32 and 83).
- [YV12] K. Yan and F.J. Verbeek, *Segmentation for high-throughput image analysis: Watershed masked clustering*, Proceedings of the 5th International Conference on Leveraging Applications of Formal Methods, Verification and Validation: Applications and Case Studies - Volume Part II (Berlin, Heidelberg), ISO/FA'12, Springer-Verlag, 2012, pp. 25–41. (cited on pages 48, 49, 80 and 81).
- [YVDvdW09] K. Yan, F. J. Verbeek, S. Le Dévédec, and B. van de Water, *Cell tracking and data analysis of in vitro tumour cells from time-lapse image sequences.*, VISAPP (1) (Alpesh Ranchordas and Helder Araújo, eds.), INSTICC Press, 2009, pp. 281–286. (cited on pages 28, 30, 80 and 81).
- [YXQZZH12] Z. Yong-Xin, L. Qing-Zhong, and P. Zhao-Hui, *Two-stage data conflict resolution based on markov logic networks*, Chinese Journal of Computers (2012). (cited on page 56).