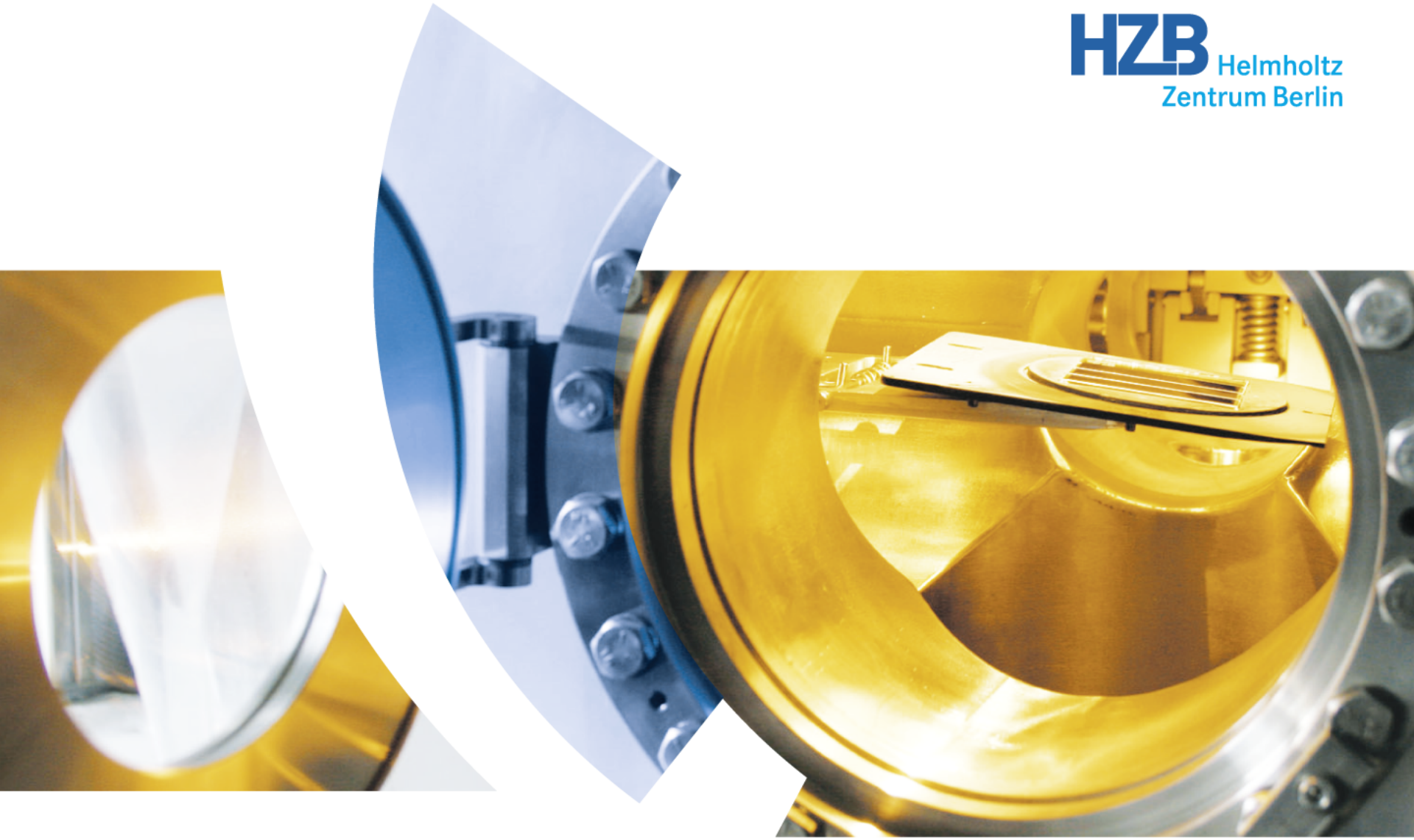
HZB Helmholtz Zentrum Berlin



FORMATS

Markus Wollgarten

Outline

- **Data formats**
  - For what?
  - What exists?
  - What is needed?
  - What's next?

HZB Helmholtz Zentrum Berlin

- Handling of **(Research) Data**

"… a set of values of subjects with respect to qualitative or quantitative variables. … Data has been described as the new oil of the digital economy." (Wikipedia)

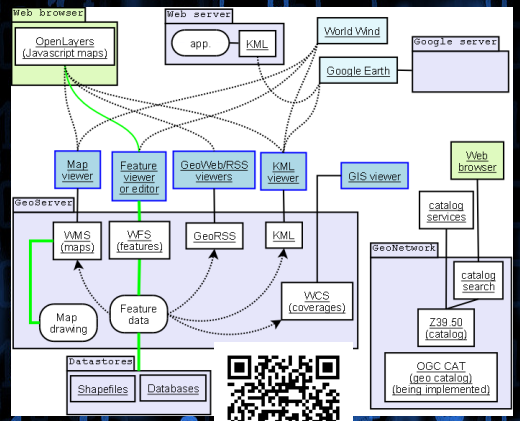
- **Given data, experimental parameters, *a priori* knowledge**
  - Time stamp, beam energy, camera configuration, lunar phase, mood of experimentalist, ...
- **Measured data, a. k. a. raw data,**
  - Image intensities, spectral intensities, sample position, ...
- **Derived data**
  - Chemical composition, size distribution, atomic structure
  - More general: *combine/correlate* measured/given data
  - Derived data can turn into "given data"

HZB Helmholtz Zentrum Berlin

- **Accessibility/Usability**
  - **Proprietary formats, many (!)**
    - some are “reverse engineered”, see e. g. “Bio Formats”, handles about 150 file formats (especially from light microscopy)

Formats – what exists? (narrowed view on microscopy)

- **Accessibility/Usability**
  - **Open formats, few (!)**
    - EMSA/MAS Standard Format for Spectral Data Exchange
    - Geographic Information System (GIS)
    - Open data network for Electron Microscopy
      - (Unsuccessfull) cost action (2016) of broad panel of research facilities (Andy Stewart, Limerick, 16 EU, 3 non-EU countries)
    - Hyperspy, not a format by itself, but correlative software tools (Python)



(wiki.gis.com)

HZB Helmholtz Zentrum Berlin

- **From EMSA/MAS white paper (N. Zalusec et al., 1991)**
  - simple and easy to use
  - both human and machine (i. e. computer) readable
  - not tied to any specific computer, programming language or operating system
  - capable of exactly presenting the data without loss of scientific content
  - contains enough information to uniquely identify the type and origin of the spectral data, to reconstruct its significance, and perform quantitative analysis

- **cont.**
  - usable with all existing electronic communication networks, telecommunications equipment (modems, Faxes) and all storage media (disks, tapes, hardcopy, print, ...)
  - supports all spectra of interest to the microanalysis community (XEDS, EELS, AES, etc.) and is flexible enough to service future data sets
  - compatible with various commercial data plotting or analysis programs
  - proposed format need not be the most efficient storage mechanism. Its primary goals, as stated above, will generally prevent storage efficiency, which is the logical role of the host system file format, not the exchange format. If anything this format will err on the side of simplicity and ease of use

(EMSA-Hyperspectral 2012)

- **From today's point of view (without priorities)**
  - Can be stored/archived/read/restored/worked-on efficiently
  - To some extent self-explanatory (markup language?)
  - Extensible
  - Certain amount of abstraction: a spectrum is a spectrum is a spectrum is a ...
  - Finds broad support by manufactures of data generating devices
  - Broad basis of users
  - ...

What's next?

- **EM@HZB**
  - Various formats in use: TIFF, DM3, EMSA, MRC, ...
- **Next steps, questions**
  - Rigid format standards/definitions – arbitrary format standards/definitions – compliance to definitions
  - One format for all? Microscopes, synchrotrons, deposition chambers, ...
  - Are there HZB specific requirements?
  - Start from zero? Jump on existing/emerging solutions like GIS, EMSA-Hyperspectral, ...?

Thank you!