# Selecting Preservation Strategies for Web Archives

**Stephan Strodl, Andreas Rauber**
Department of Software Technology
and Interactive Systems
Vienna University of Technology

# Motivation

- web archive systems store enormous amount of data
- no guarantee to reopen in 5, 10 or 20 years

→ useless, waste of time & money?

- digital preservation

- special challenges of web archives
  - amount of data
  - heterogeneity of file formats
  - quality of data (wrong mime type)
  - crawler specific characteristics of data collection

# Motivation

- different strategies for preservation of web archives
  - original
  - migration (ASCII, picture, video clip)
  - standardization (minimal HTML)

- how do you know what is most suitable for your needs?
- what are your requirements?
- how do you measure and evaluate the results of the preservation strategies?

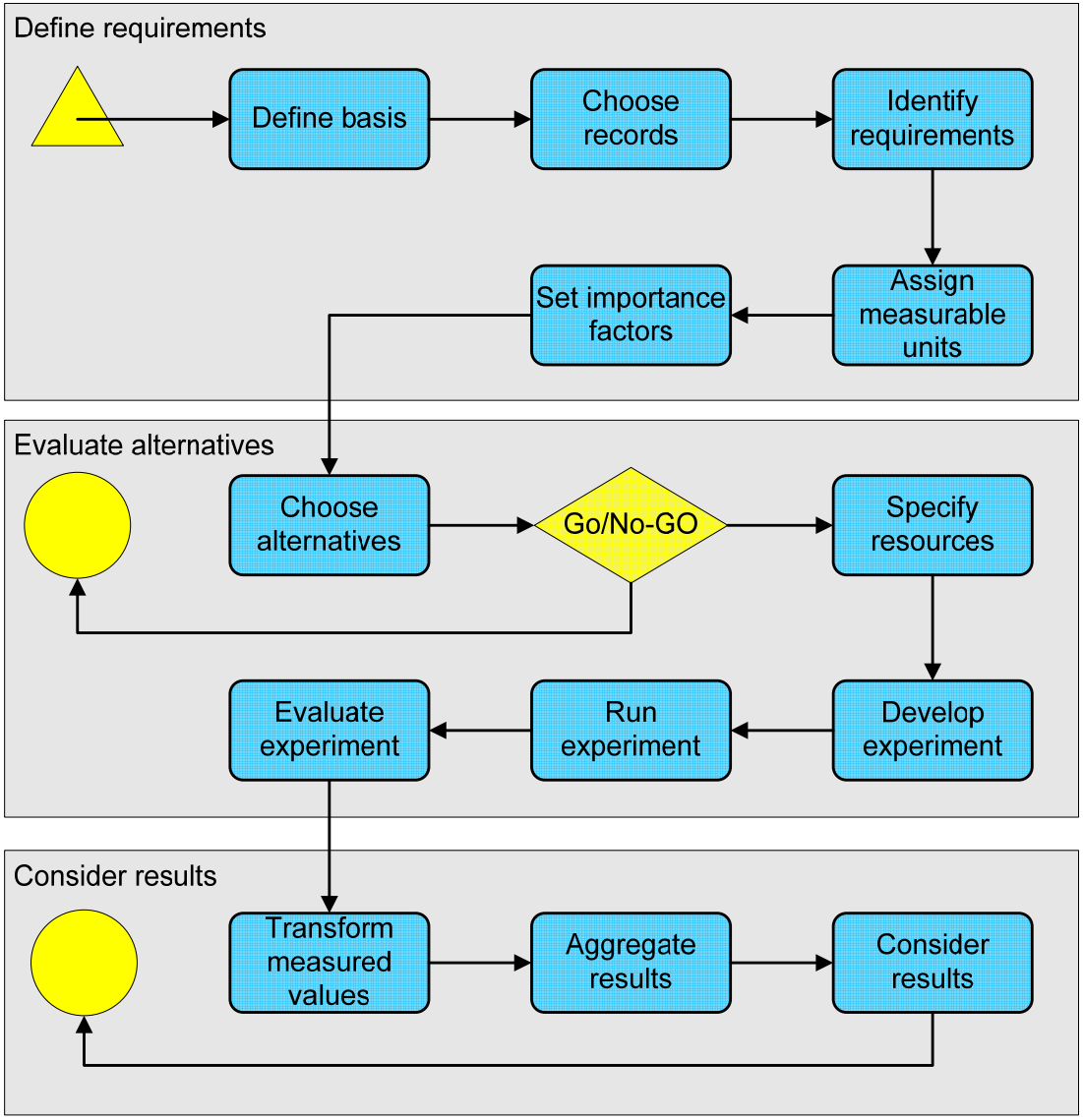
# Goals

- motivate and allow operators of web archives to precisely specify their preservation requirements (future usage of web archive)
- provide structured model to describe and document these
- create defined setting to evaluate preservation strategies
- document outcome of evaluations to allow informed, accountable decision

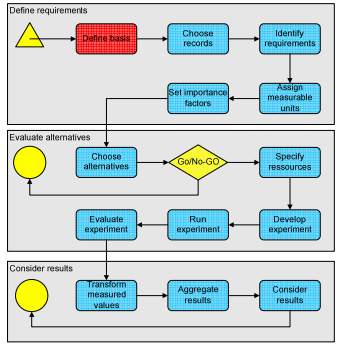
# Utility Analysis

- cost-benefit analysis model
- used in the infrastructure sector
- adapted for digital preservation needs
- 14 steps grouped into 3 phases
- framework in cooperation of Vienna University of Technology and National Archive Netherlands

# Process Overview

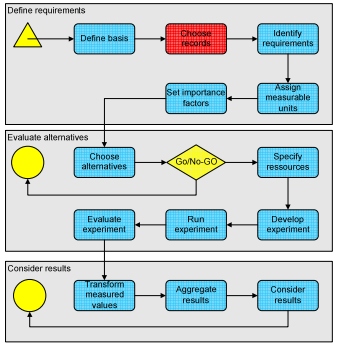


# Define basis



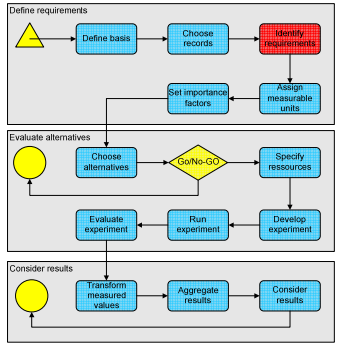

- types of records (e.g. Java applets, audio streams, Flash, ..)
- what are the essential characteristics?
  - content, context(!), structure, form and behaviour
- specific task of web archives (e.g. e-gov vs. historic websites)
- requirements
  - metadata
  - authenticity, reliability, integrity, usability

# Choose objects/records



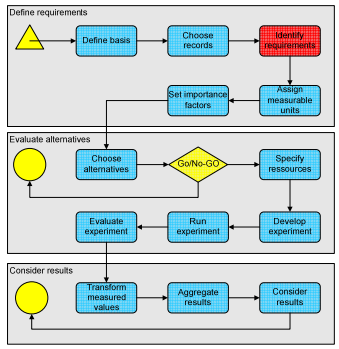

- choose sample records
  - a test-bed repository
  - from own collection
- choice of records affects the evaluation

# Identify objectives (1)




- list all requirements and goals in tree structure
- start from high-level goals
- break down to fine-granular, specific criteria

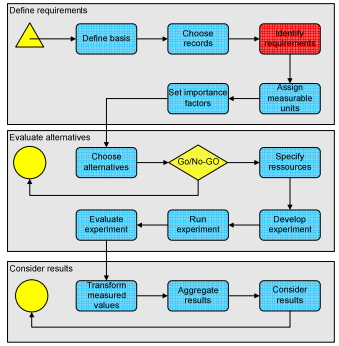
# Identify objectives (2)




- usually 4 top-level branches:
  - object characteristics *(content, metadata ...)*
  - record characteristics *(context, relations, ...)*
  - process characteristics *(scalability, error detection, ...)*
  - costs *(set-up, per object, HW/SW, personnel, ...)*

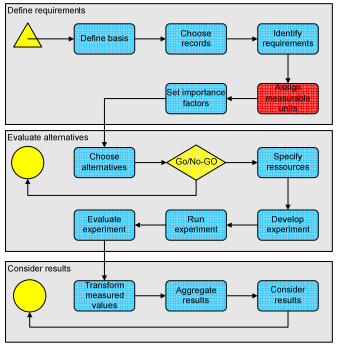
- define requirements for web archives
  - preserve picture, video clip, text content, interactivity
  - search, links, metadata

# Identify objectives (3)



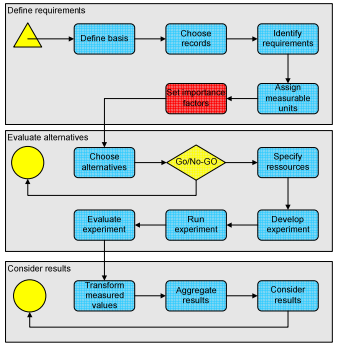

- objective tree with several hundred leaves
- usually created in workshops, brainstorming sessions
- re-using branches from similar institutions, collection holdings, ...

# Assign measurable units



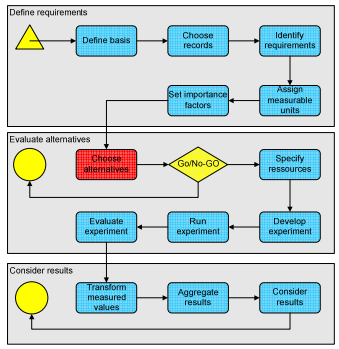

- ensure that leaf criteria are objectively (and automatically) measurable
  - seconds/Euro per object
  - bits color depth
  - ...
- subjective scales where necessary
  - diffusion of file format
  - amount of (expected) support
  - ...

# Set importance factors




- set importance factors
- not all leaf criteria are equally important
- set relative importance of all siblings in a branch
- weights are propagated down the tree to the leaves

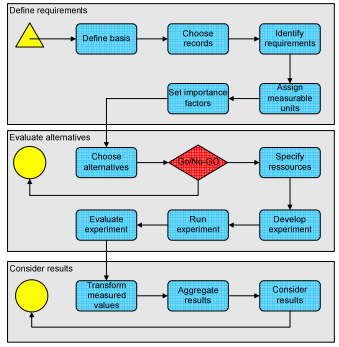
# Choose alternatives




- list and formally describe the preservation action possibilities to be evaluated
  - tool, version
  - operating system
  - parameters

- alternatives for web archives
  - original
  - migration (ASCII, picture, video clip)
  - standardization (minimal HTML)

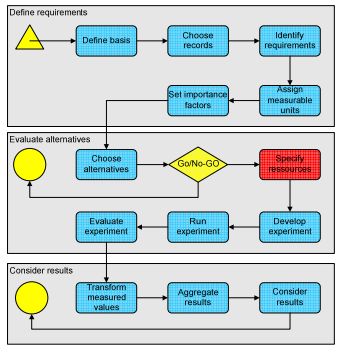
# Go/No-Go




- deliberate step for taking a decision whether it will be useful and cost-effective to continue the procedure, given
  - the resources to be spent (people, money)
  - the expected result(s).

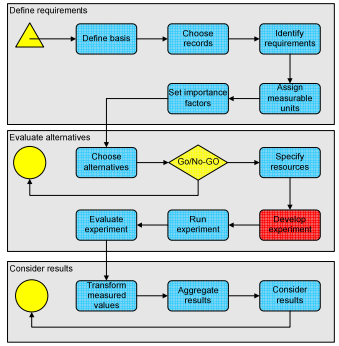
- review of the experiment/ evaluation process design so far
  - e.g. is the design correct and optimal?
  - is the design complete (given the objectives).

# Specify resources



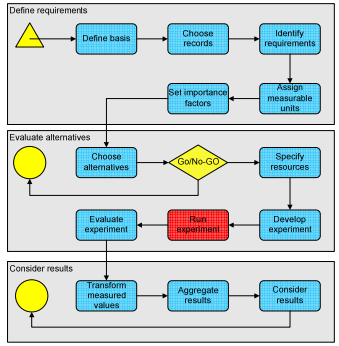

- detailed design and overview of the resources
  - human resources (qualification, roles, responsibility, …)
  - technical requirements (hardware and software components)
  - time (time to run experiment,...)
  - cost (costs of the experiments,...)

# Develop experiment



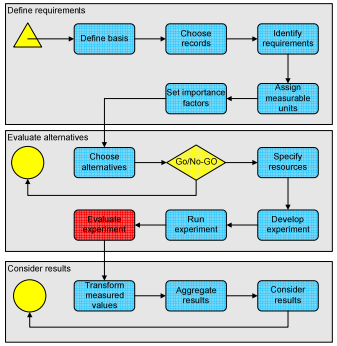

- formulate for each experiment a detailed plan
  - includes builds build and test software components
  - mechanism to capture the result
  - workflow/sequence of activities

# Run experiment



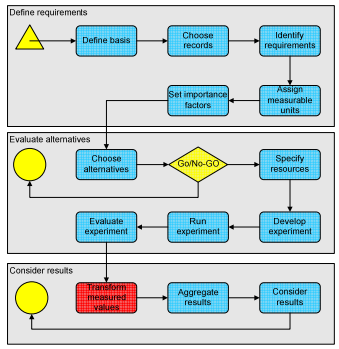

- run experiment with the previously defined sample records
- the whole process need to be documented
- e.g. convert html file to pdf

# Evaluate experiment



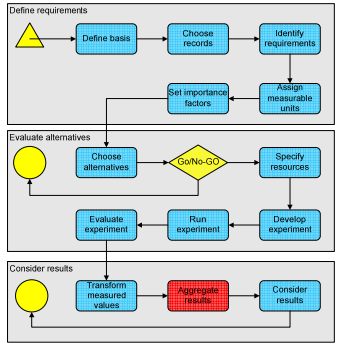

- evaluate how successfully the requirements are met
- measure performance with respect to leaf criteria in the objective tree
- document the results

# Transform measured values



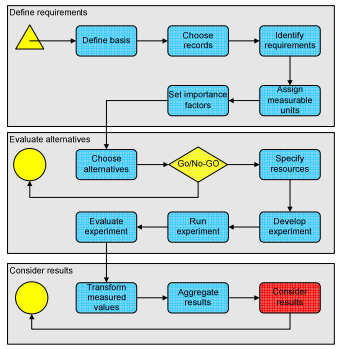

- measures come in seconds, euro, bits, goodness values,…
- need to make them comparable
- transform measured values to uniform scale
- transformation tables for each leaf criterion
- linear transformation, logarithmic, special scale
- scale 1-5 plus "not-acceptable"

# Aggregate values



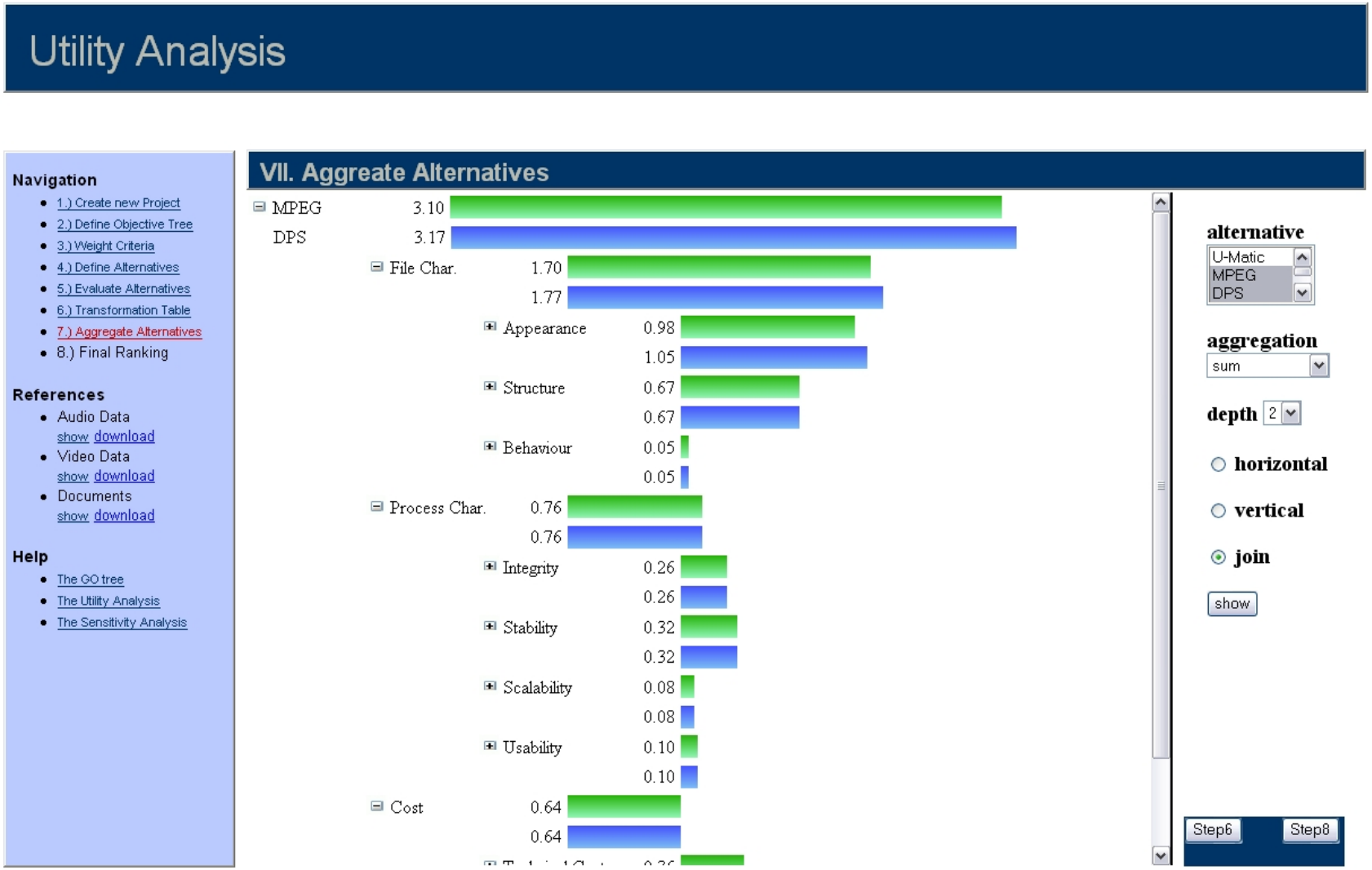

- multiply the transformed measured values in the leaf nodes with the leaf weights
- sum up the transformed weighted values over all branches of the tree
- creates performance values for each alternative on each of the sub-criteria identified

# Consider results




- rank alternatives according to overall utility value at root
- performance of each alternative
  - overall
  - for each sub-criterion (branch)
- allows performance measurement of combinations of strategies
- final sensitivity analysis against minor fluctuations in
  - measured values
  - importance factors

# Digital Pres. Utility Analysis Tool



# Benefits

- a simple, methodologically sound model to specify and document requirements
- repeatable and documented evaluation for informed and accountable decisions
- set of templates to assist institutions
- generic workflow that can easily be integrated in different institutional settings

# Conclusion

- important to consider preservation for web archives
- web archive suitable for combination of strategies
- need a profound knowledge of future use of web archives

# Questions ?