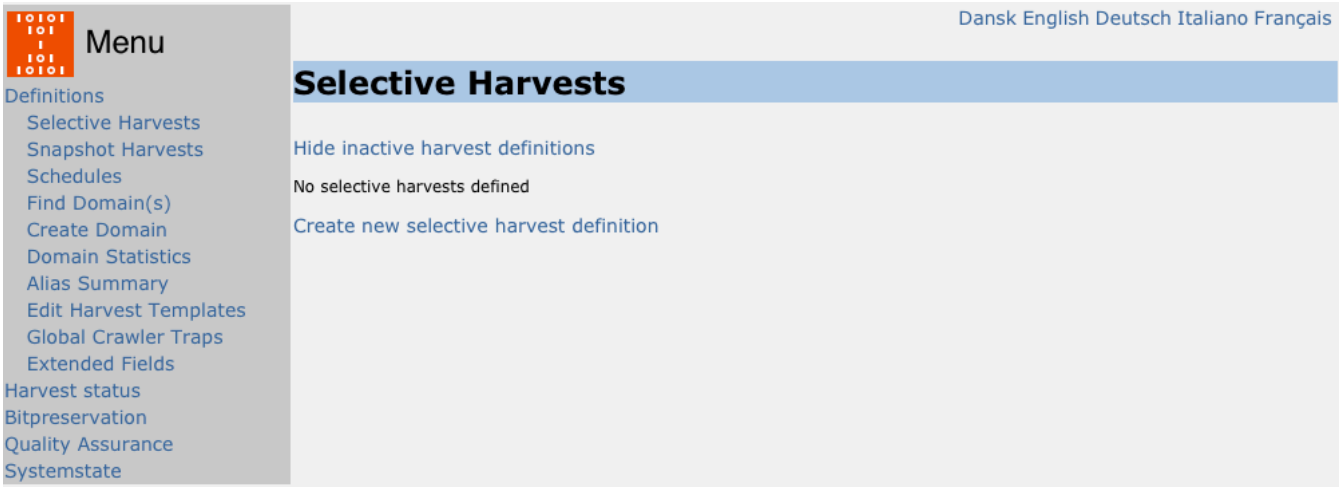
Selective and Event Harvests

- Creating/editing a selective harvest
    - Easy creation of non existing domains
- Event harvest
- Adding seeds to an event harvest

A Selective Harvest consists of one or more harvest configurations (possibly from different domains) and a schedule (e.g. once-per-day). In NetarchiveSuite, Event Harvests are treated as a special kind of Selective Harvest. An Event Harvest is generated directly from a set of seeds, a Heritrix crawler-bean template, and a schedule.

The front page by default shows the list of selective harvests:



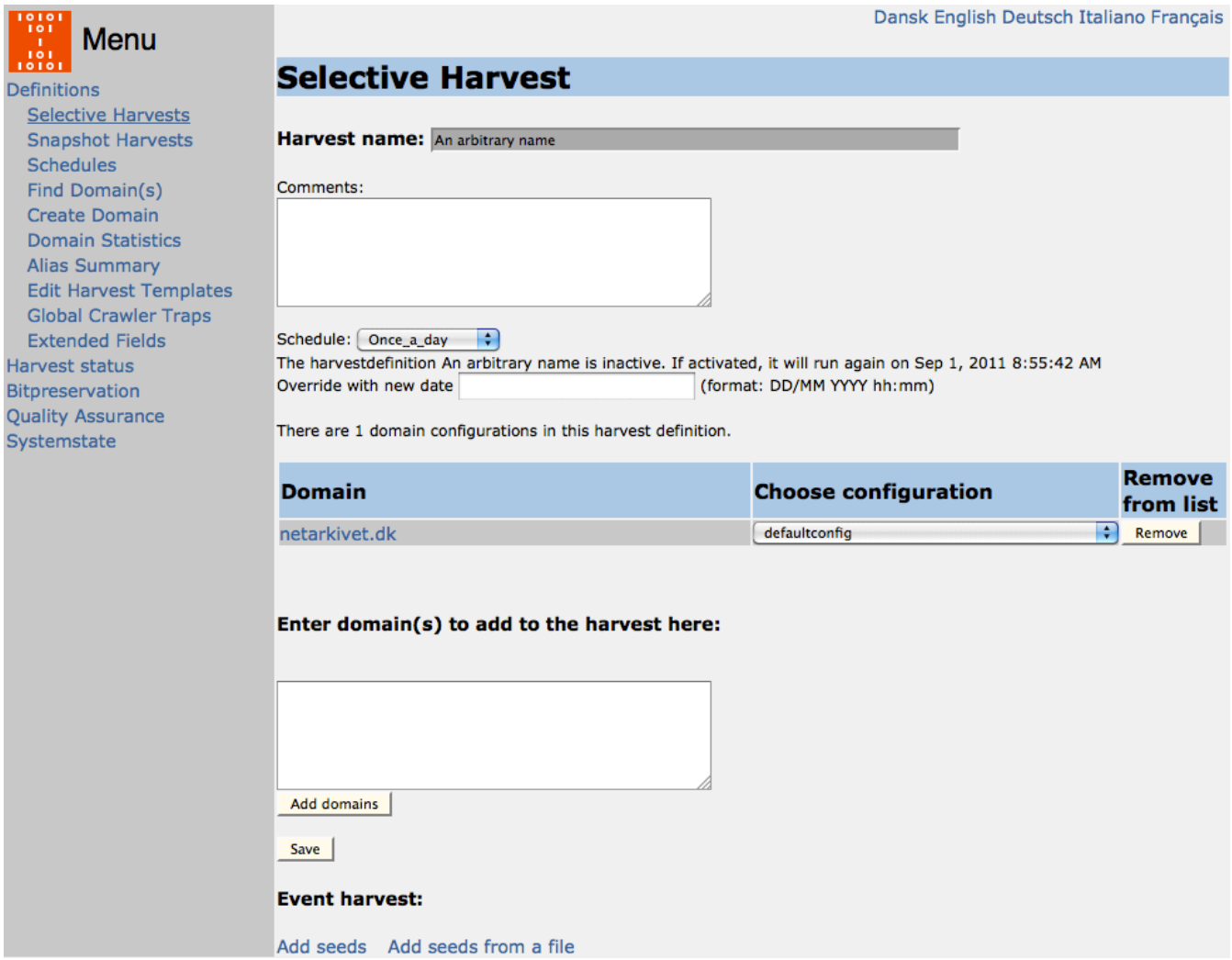
You can *Activate* an inactive harvest definition and *Deactivate* an active harvest definition. If you deactivate a running harvest, the system will finish the running jobs.

You can hide inactive harvest definitions to avoid screen clutter.

Click on **Edit** to change an existing harvest definition.

Click on **History** if you wish to trace back all completed jobs from a given harvest definition.

Creating/editing a selective harvest



Create a new selective harvest definition by pressing **Create new selective harvestdefinition** from the frontpage.

Give the harvestdefinition a recognizable harvest name – you cannot change it or delete it later. If necessary add a comment.

Choose a schedule from the dropdown list (see the relevant section for how to define new schedules).

Now you can add domains to the harvestdefinition.

Write the name of the domains you want to add in the box **Enter domain**(s) to add to the harvest here and click on **Add domains**.

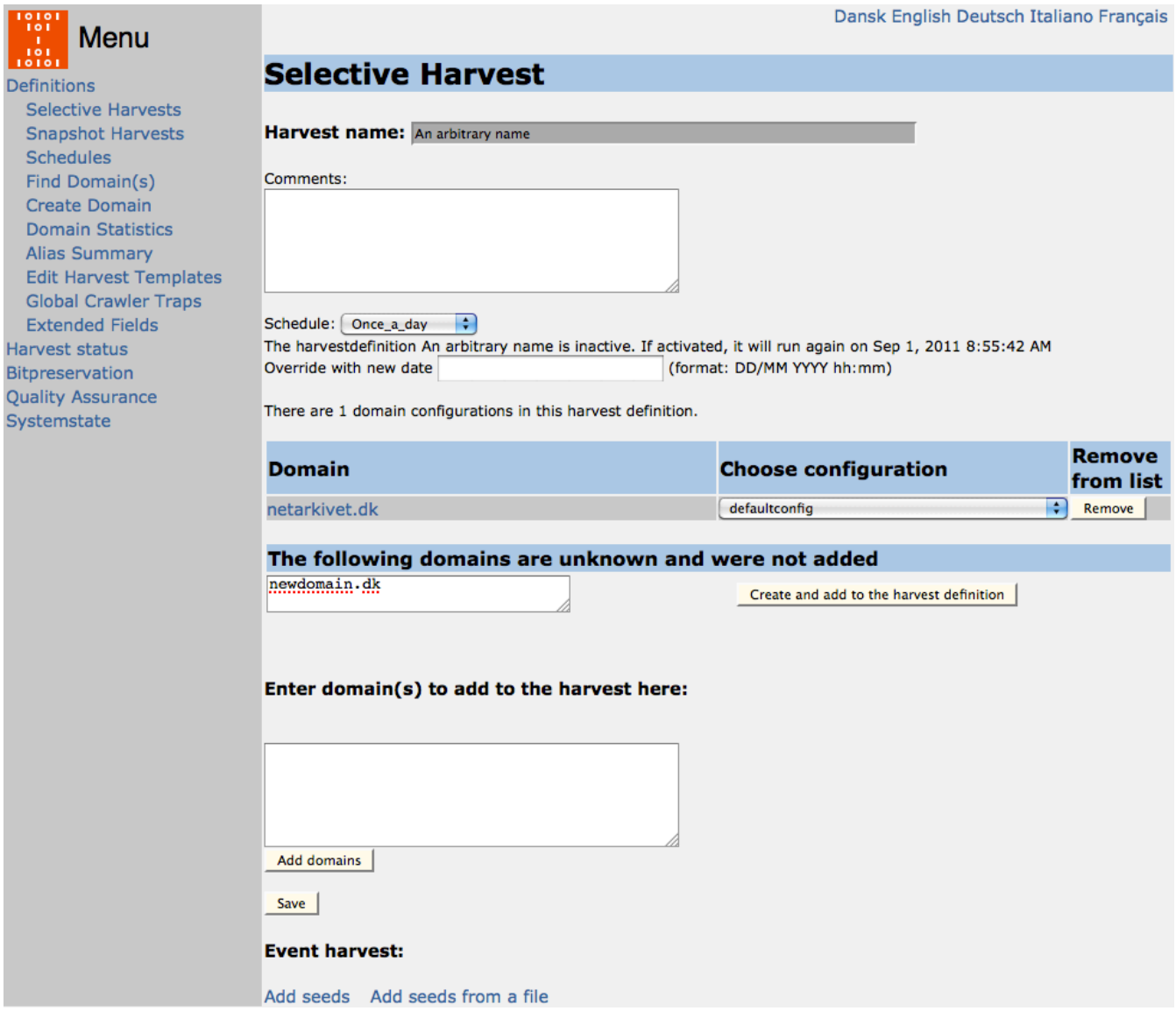
The added domains will appear in the column **Domain**.

For each added domain, choose the wanted configuration from the dropdown list for each domain. Press **Save** to save the harvestdefinition.

The scheduling of selective harvest definitions can be overridden by filling out the input field *Override with new date*. Simply set the date to whenever you wish the harvest definition to run next time. The scheduling of the harvest definition will continue from that point in time.

Note that **newly created harvests are always inactive**. You must return to the front page and activate the harvest if you want it to run.

Easy creation of non existing domains



When adding a domain that does not already exist you will see the warning **The following domains are unknown and were not added**. You can simply add the unknown domains to the database and your harvestdefinition by clicking **Create and add to the harvestdefinition**.

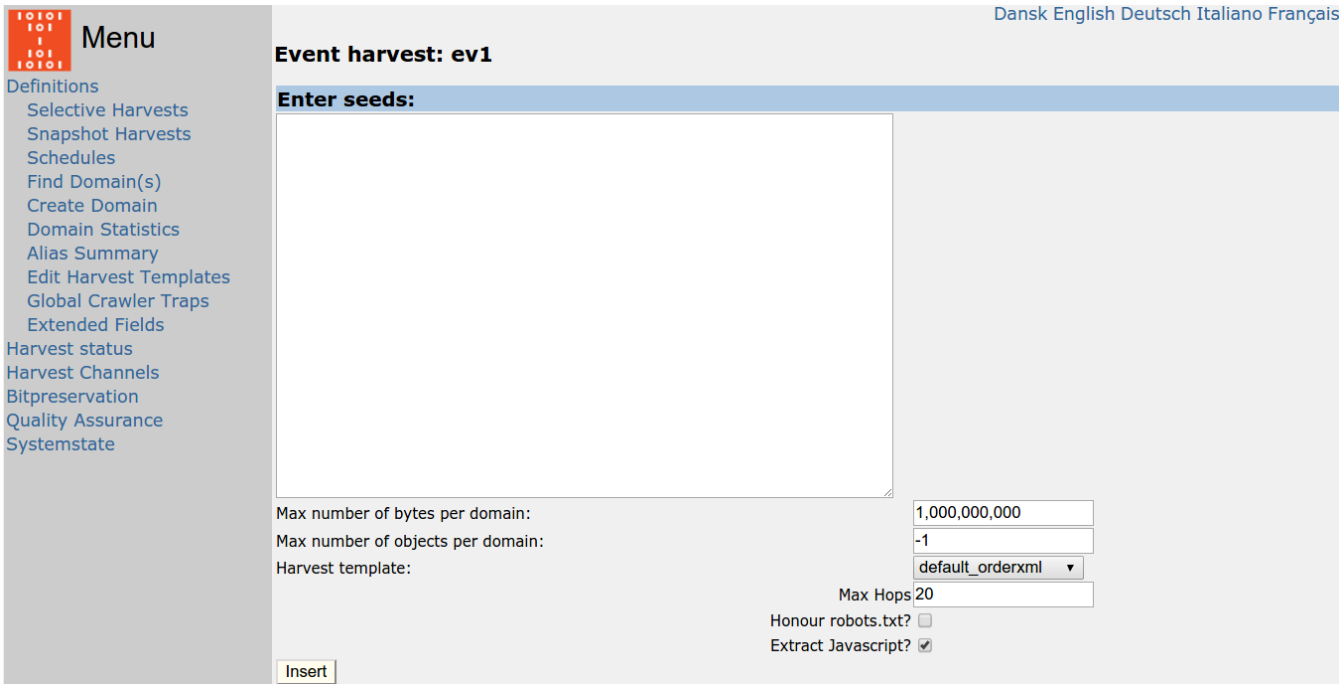
Event harvest

An Event Harvest typically represents a response to a "Real World" event where one wishes to create a special harvest capturing the web's response to the happening. Typically, as a curator, one gathers links to many relevant pages and sites (perhaps hundreds) which one wishes to harvest in a single coherent harvesting process. The Event Harvest functionality in NetarchiveSuite makes this easy to do.

Event harvests are treated almost the same as selective harvests in the system. The difference lies only in the additional functionality for adding and configuring many domains and seeds at once. This allows the operator to add a large number of seed URLs without having to edit configurations and seedlists for all the affected domains by hand.

To start, simply create and save a Selective Harvest as usual but don't add any domains. You may want to include something in the harvest name to make clear that it is an Event Harvest e.g. "Event Harvest: 2022 General Election".

Adding seeds to an event harvest



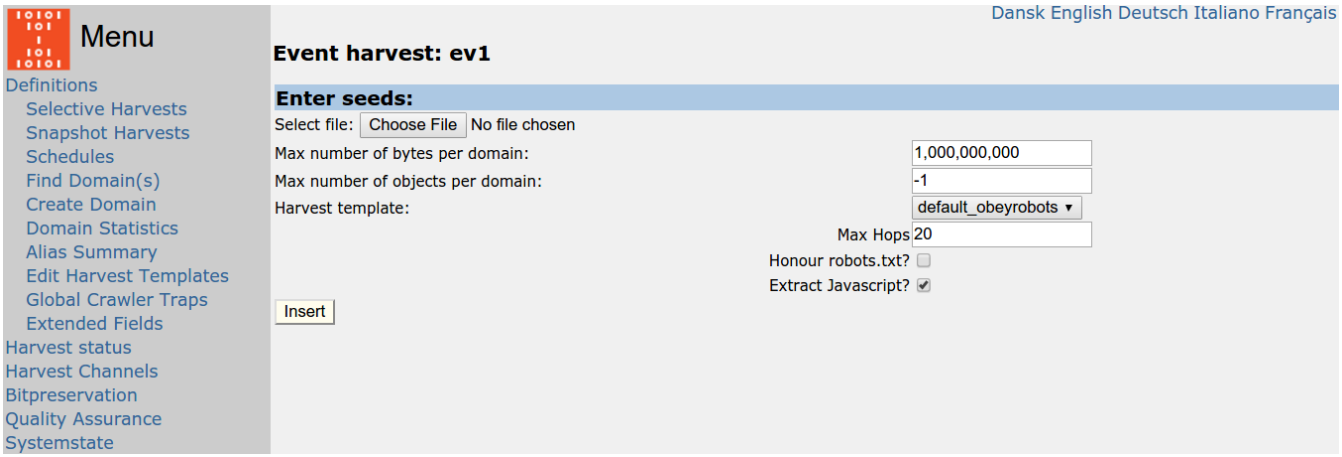
Click on the Add seeds link at the bottom of the *Selective Harvest* page. Enter identified start-URLs covering the event in the **Enter seeds:** box. In **Max number of bytes per domain** enter your preference, and choose a .cxml template for the harvest. Specify also the required depth of the harvest (MAX_HOPS), whether it should honout robots.txt, and whether or not Heritrix should attempt to extract links from Javascript elements.

All seeds added together will use the same template, so to harvest different seeds with different templates you need to add them in groups, one group per desired template.

Pressing **Insert** starts the power-adding function. This function runs through the entered seeds one by one and does the following with each seed:

1. Finds the domain from which the seed derives, or creates it if necessary
2. Creates a seedlist whose name contains the name of the harvestdefinition and the template (and the specified maximum bytes and/or objects)
3. Creates a configuration with the same name as the above seedlist and selects that seedlist for use with the configuration. (If the seedlist to create in (2) or the configuration to create in (3) already exist the system will only add the new URLs to the existing seedlist.)
4. Adds that harvest configuration to the event harvest

You can also use *Add seeds from a file*. This allows you upload a file with the seeds instead of entering the seeds in a text field. Otherwise the functionality is the same.



Don't forget to **activate the Event Harvest** if you actually want it to run.